Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in

Statistica per le Tecnologie e le Scienze

Clustering non parametrico per la separazione efficiente di sorgenti di raggi gamma dal *background* diffuso

Relatore: Prof.ssa Giovanna Menardi
Dipartimento di Scienze Statistiche

Laureando: Francesco Freni
Matricola n. 2001862

Anno Accademico 2022/2023

Indice

Introduzione

L'universo è teatro di reazioni fisiche spesso collegate alla produzione di grandi moli di energia, che si manifesta in emissioni di raggi gamma, ovvero forme di radiazioni elettromagnetiche ad elevato quantitativo energetico. Le radiazioni elettromagnetiche sono le principali fonti di informazione astronomica, in quanto tengono traccia della natura degli eventi cosmici che le originano, ad esempio buchi neri supermassicci, stelle di neutroni che si fondono e flussi di gas caldo. Dunque, lo studio di raggi gamma permette una più profonda comprensione dell'ambiente astrofisico, in quanto aiuta ad indagare i meccanismi che descrivono la creazione e l'accelerazione delle particelle emesse dai corpi celesti.

Il telescopio spaziale per raggi gamma *Fermi*[1] è un osservatorio spaziale internazionale approvato nel 2001 dalla NASA ed il cui lancio è avvenuto nell'estate del 2008. L'osservatorio è stato istituito con lo scopo, tra gli altri, di identificare le sorgenti di raggi gamma non note e permettere la loro discriminazione dalla radiazione gamma diffusa.

Esso rileva i fotoni, ovvero particelle prive di massa, costituenti principali delle radiazioni elettromagnetiche. Il telescopio Fermi è sensibile alle onde elettromagnetiche emesse da corpi celesti nell'intervallo di energie tra 8 mila elettronvolt (8 keV) e 300 miliardi di elettronvolt (300 GeV), ovvero quantitativi energetici molto superiori alla luce visibile ad occhio nudo.

[1] `https://www.nasa.gov/content/fermi/overview/`

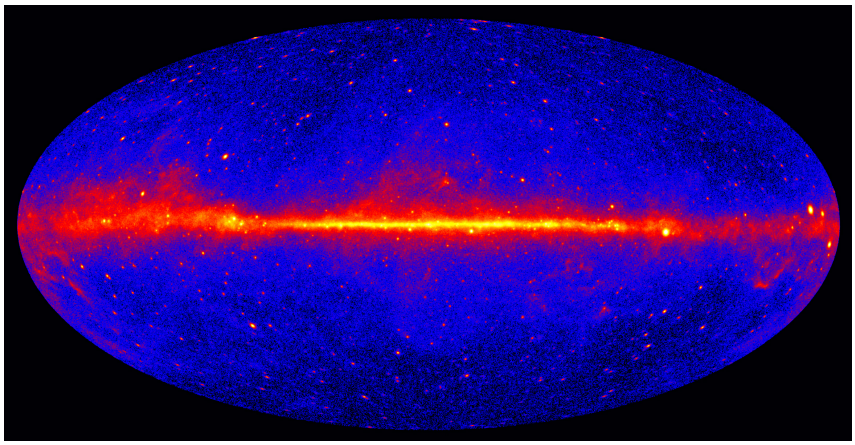

Figura 1: Rappresentazione dei conteggi dei fotoni ad elevata energia rilevati dal telescopio Fermi utilizzando le emissioni rilevate nell'arco di cinque anni. L'energia riportata è superiore a 1 GeV.

In particolare, il Fermi comprende due strumenti: il *Large Area Telescope* (LAT) e il *Gamma-Ray Burst Monitor* (GBM). Il primo osserva tutta la volta celeste ogni tre ore e permette di cogliere radiazioni gamma comprese tra 20 MeV e 300 GeV, con picchi massimi fino a 2 TeV, mentre il secondo è utilizzato per lo studio di esplosioni di raggi gamma e rileva radiazioni comprese tra 8 keV e 40 MeV. In questo elaborato si fa riferimento ai dati raccolti dal Fermi LAT.

Nella Figura 1 è riportata l'immagine della mappa galattica, dove è rappresentata l'emissione di energia rilevata dal telescopio Fermi a seguito della missione che ha raccolto dati lungo un periodo di cinque anni. L'immagine rappresenta la mappa dell'universo sfruttando le coordinate galattiche, ossia latitudine e longitudine galattica, e la luminosità di ogni porzione di mappa è proporzionale all'energia emessa in quella data zona. Nella regione lungo il piano galattico è presente la galassia a cui appartiene il nostro sistema solare, la Via Lattea, al cui centro è presente un buco nero supermassiccio e la cui luminosità indica la presenza di numerosi eventi ad alta energia.

Il dato di cui si dispone, il fotone, è direzionale, ovvero un vettore

sulla sfera. L'obiettivo che ha condotto all'ideazione del telescopio Fermi è lo studio della direzione di emissione dei fotoni, per cui si utilizza il sistema di coordinate cartesiane sulla sfera di raggio unitario.

La luce rilevata in una regione dello spazio è il risultato della sovrapposizione di differenti sorgenti fisiche altamente eterogenee, sia in dimensione sia in variabilità, che si possono trovare lungo il piano galattico (sorgenti galattiche) o lontane da questo (sorgenti extragalattiche). Una distinzione importante è quella tra sorgenti localizzate e diffuse. Le prime appaiono come regioni ad alta intensità nitidamente definite, mentre le seconde appaiono come regioni estese di intensità variabile in una porzione dello spazio, ad esempio le nuvole di gas nella Via Lattea, che emettono fotoni quando raggi cosmici collidono con elio ed idrogeno contenuti in esse.

Oltre alle sorgenti, è presente il rumore di fondo (*background*), composto da emissioni galattiche diffuse e radiazioni isotropiche diffuse. Il termine diffuso si riferisce a una distribuzione di raggi gamma con una componente maggiore presso il centro galattico (latitudine 0°) che diminuisce allontanandosi da questo. Il termine isotropico si riferisce a una distribuzione uniforme su tutta la sfera. Le direzioni presso il centro galattico, per esempio, presentano un alto livello di emissione da nuvole di gas diffuso, mentre l'emissione a latitudini galattiche elevate è dominata da una componente di fondo isotropica. In questa seconda regione risulta generalmente più semplice identificare sorgenti, ovvero quantificarne l'evidenza e localizzarle in una regione dello spazio.

L'obiettivo di questo elaborato è quello di risalire alla direzione di emissione dei fotoni, distinguendo i raggi gamma emessi dalle sorgenti celesti da quelli emessi dal rumore di fondo (*background*). Tale obiettivo è stato perseguito in letteratura seguendo due approcci principali: uno basato sull'individuazione di un'unica sorgente e l'altro, più recente, basato sull'individuazione di più sorgenti. Il primo ap-

proccio (Hobson, 2009, par. 7.4) richiede la suddivisione della sfera celeste in piccole regioni e la possibile presenza di una nuova sorgente viene valutata per ogni pixel. In particolare, si assume che il conteggio dei fotoni in ogni pixel segua una distribuzione di Poisson e la significatività di ogni sorgente viene valutata grazie a test basati sul log-rapporto di verosimiglianza (Mattox *et al.*, 1996). Il secondo approccio, invece, consiste nell'identificare e localizzare più sorgenti simultaneamente, senza suddividere la sfera in regioni più piccole (Hobson, 2009, par. 7.3) ed esempi recenti sono basati su modelli mistura.

In questo elaborato, similmente a Montin *et al.* (2023), il problema è affrontato grazie al *clustering* non parametrico, efficacemente adattato al contesto in esame. Lavorare su tutta la sfera comporta complicazioni dovute all'elevata concentrazione dei fotoni e all'enorme mole di dati. Nell'elaborato queste vengono superate grazie all'utilizzo di una griglia sferica opportunamente partizionata in modo da non suddividere sorgenti. Lo studio dei *cluster* verrà effettuato sulle regioni individuate separate dalle celle vuote.

La trattazione si sviluppa come segue: nel Capitolo 1 viene approfondito il tema del *clustering* non parametrico in un contesto generale che prescinde dalle applicazioni in ambito astrofisico. Nel Capitolo 2 viene proposta un'estensione del *clustering* modale per l'identificazione dei raggi gamma, che tenga conto dell'elevata concentrazione dei fotoni e della mole dei dati, cercando di superare le difficoltà che emergono dal punto di vista computazionale. Nel Capitolo 3 viene illustrata un'applicazione della procedura proposta ad un insieme di dati simulati.

Capitolo 1

Clustering non parametrico

1.1 Introduzione

Come evidenziato nell'introduzione, il problema in esame può essere ricondotto ad un contesto di analisi di raggruppamento (*cluster analysis*). Le tecniche esistenti progettate per questo scopo, applicabili sia come analisi preliminari, sia come obiettivo conclusivo delle analisi, possono essere suddivise in due tipologie.

La prima è composta dai metodi gerarchici e dai metodi di partizione, che si fondano sul concetto di distanza o dissimilarità. L'idea alla base della formazione dei gruppi consiste nel cercare di minimizzare la distanza tra le osservazioni appartenenti ad uno stesso *cluster* e di massimizzare la distanza tra le osservazioni appartenenti a gruppi differenti. Nei metodi di partizione, fra i quali ricade il metodo delle *k-medie*, si vanno dunque a ricercare delle suddivisioni dello spazio campionario e si sceglie quella che minimizza una qualche funzione obiettivo. Il *k-means*, ad esempio, parte da una partizione iniziale dei dati in k gruppi e, dopo aver calcolato i centroidi di ogni *cluster*, ovvero le medie delle osservazioni appartenenti ai vari gruppi, assegna ogni punto al *cluster* il cui centroide presenta distanza minima da esso. La procedura prosegue con l'aggiornamento delle medie dei gruppi e viene ripetuta iterativamente fino a convergenza. I metodi gerarchici, invece, valutano successivamente delle partizioni nidifica-

te sulla base della distanza tra punti e non costruiscono una singola partizione con *k cluster*, perché considerano ogni valore di *k*.

Nonostante la semplicità concettuale e computazionale, questi metodi non presentano fondamenti statistici alla base, poiché si basano su una definizione euristica di gruppo, non permettono l'utilizzo di procedure inferenziali per la valutazione dei risultati e la definizione del numero di gruppi presenti nei dati. Per un approfondimento si veda Kaufman e Rousseeuw (2005).

Il secondo approccio al *clustering* comprende i metodi basati sulla densità, per cui i gruppi vengono associati a delle caratteristiche specifiche della distribuzione di probabilità che si assume essere sottostante ai dati. Queste tecniche hanno portato ad un approccio statisticamente più rigoroso, in quanto permettono di fare inferenza sul numero dei gruppi e sulla bontà della partizione. Questa idea è stata sviluppata in due direzioni: da un lato il *clustering* basato su modelli parametrici, dall'altro il *clustering* modale o non parametrico.

La formulazione parametrica (*model-based approach*) assume che la densità dei dati sia descritta da un modello a mistura finita, in cui ogni gruppo è associato ad una componente, tipicamente appartenente ad una prefissata famiglia parametrica. Il problema di *clustering* si riconduce pertanto ad un classico problema di inferenza, il cui obiettivo è stimare i parametri del modello specificato. L'approccio di stima privilegiato in questo contesto è basato sulla massima verosimiglianza. Questo metodo presenta indubbi vantaggi legati all'uso di un modello parametrico, come la parsimonia, la presenza di procedure di stima consolidate ed utilizzabili anche in presenza di dati complessi e l'interpretabilità. D'altra parte, poiché si assume che i gruppi abbiano una forma compatibile con il modello parametrico scelto, nel caso in cui l'assunzione parametrica alla base venisse violata, la qualità dei risultati sarebbe compromessa. Per una rassegna sul *clustering* parametrico si veda Bouveyron *et al.* (2019).

L'approccio non parametrico, come suggerisce il nome, fa ricor-

so a metodi non parametrici per la stima della densità sottostante ai dati, ma si contraddistingue anche per una differente nozione di *cluster*. Introdotto informalmente da Carmichael e Julius (1968), ha alle proprie basi una definizione intuitiva di gruppi, legati a regioni dello spazio relativamente densamente popolate, circondate da regioni relativamente vuote. Da un punto di vista probabilistico, questo si traduce nell'associare ogni *cluster* ad una regione ad elevata densità del supporto, ovvero ad una *regione modale* o *dominio di attrazione* di una moda. Questa nozione risulta conforme alla definizione di sorgente quale picco di concentrazione di fotoni che emerge dal *background* e per questo viene adottata nel presente elaborato. L'idea ha trovato una sua formalizzazione in Chacón (2015), grazie alla teoria Morse, una branca della geometria differenziale che studia la topologia dello spazio in cui una funzione è definita tramite l'analisi dei suoi punti critici. Un *cluster* è un involucro instabile del flusso negativo del gradiente della funzione di densità sottostante i dati associato ai punti di massimo. Il concetto può essere più intuitivamente definito: se si associasse la densità ad una catena montuosa i cui picchi rappresentano le mode, il dominio di attrazione di una moda sarebbe identificato dalla regione bagnata da un flusso d'acqua, fatto scorrere in assenza di attrito, fuoriuscente dalla relativa cima. Si veda la Figura 1.1 per un'illustrazione. Oltre al vantaggio di definire formalmente il contesto di lavoro, che consente un approccio inferenziale al problema di *clustering*, la formulazione non parametrica fornisce libertà alla modellazione della forma dei singoli gruppi, pur mantenendo rigore dal punto di vista statistico. Inoltre, il numero di gruppi è una proprietà intrinseca del meccanismo generatore dei dati e la sua determinazione è parte della procedura di stima. Operativamente, le questioni critiche da affrontare riguardano la stima della funzione di densità e la ricerca delle sue regioni modali, approfondite nel paragrafo che segue.

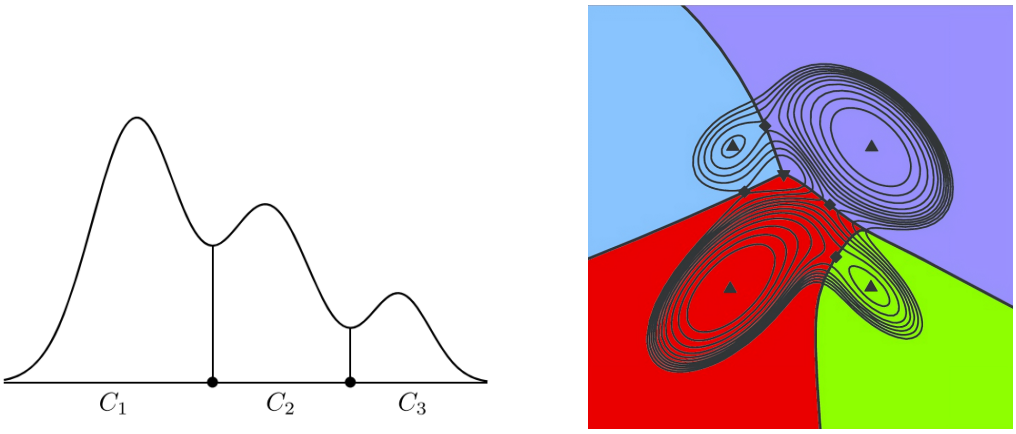

Figura 1.1: Un esempio di densità univariata (a sinistra) e bivariata (a destra), con i relativi domini di attrazione delle mode. Le immagini sono tratte da Chacón (2015).

1.2 Stima non parametrica della densità

1.2.1 Aspetti generali

Tra gli stimatori non parametrici di una funzione di densità, il metodo del nucleo (Parzen, 1962) è il più noto e diffuso, per ragioni di convenienza matematica e operativa. Questo metodo rappresenta una generalizzazione del concetto di istogramma e permette la stima della funzione di densità evitando il ricorso a forme parametriche (Figura 1.2).

Nel caso univariato, dato un campione $\mathscr{X} = (x_1, ..., x_n)$ di n realizzazioni da una funzione di densità $f : \mathbb{R} \to \mathbb{R}^+ \cup \{0\}$, lo stimatore *kernel* è così definito:

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - x_i) \tag{1.1}$$

dove $K_h(x) = \frac{1}{n} K(\frac{x}{h})$, $K(\cdot)$ è il *kernel*, funzione non negativa e che integra a uno e $h > 0$ è una costante di lisciamento. La scelta di K determina la forma della funzione, ma si è visto in pratica che differenti specificazioni del *kernel* non portano a cambiamenti sostanziali della stima di densità (Chacón e Duong, 2018). Al contrario, il parametro

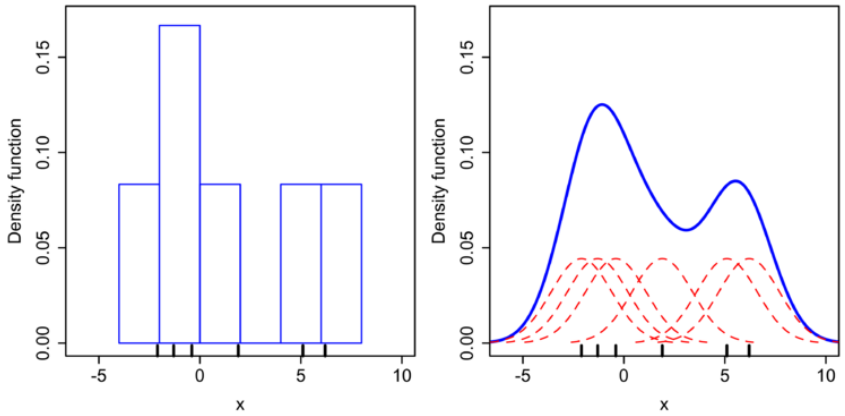

Figura 1.2: Confronto fra l'istogramma (a sinistra) e la stima della densità *kernel* costruiti utilizzando gli stessi dati (a destra).

di lisciamento h deve essere adeguatamente scelto, poiché determina l'ampiezza di banda con cui si includono le osservazioni per la stima della densità in ogni punto: valori grandi renderanno liscia $\hat{f}$, con il conseguente rischio di eliminare delle mode, mentre valori piccoli porteranno ad una curva frastagliata e sovra-adattata ai dati, aumentando così la possibilità di includere mode spurie.

Analogamente, nel caso multivariato, dato un campione $\mathscr{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$ di n realizzazioni da un vettore casuale d-dimensionale da una funzione di densità $f : \mathbb{R}^d \to \mathbb{R}^+ \cup \{0\}$, lo stimatore *kernel* di tale densità è definito come segue:

$$\hat{f}_H(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} K_H(\boldsymbol{x} - \boldsymbol{x}_i) \tag{1.2}$$

dove $K_H(\boldsymbol{x}) = |H|^{-1/2} K(H^{-1/2}\boldsymbol{x})$ e $K(\cdot)$ è una funzione *kernel* d-variata, simmetrica e che integra a uno. Il parametro di lisciamento, H, è una matrice quadrata di dimensione d, simmetrica e definita positiva. In questo caso la selezione del parametro di lisciamento comporta la determinazione di *d(d+1)/2* parametri, e, di conseguenza, un maggior guadagno di flessibilità a scapito di una maggiore complessità della scelta. Una semplificazione consiste nell'adottare come parametro di lisciamento una matrice diagonale con elementi diagonali

pari ad h, ponendo $H = h^2 I$, $h > 0$. Ciò permette di ricondursi ad una situazione matematicamente più semplice:

$$\hat{f}_h(\boldsymbol{x}) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{\boldsymbol{x} - \boldsymbol{x}_i}{h}\right). \tag{1.3}$$

1.2.2 Il parametro di lisciamento

Criteri di selezione

Data la criticità relativa alla selezione del parametro di lisciamento, la letteratura è ricca di contributi volti a determinare criteri di ottimalità. Alla base di tali criteri vi è, tipicamente, la valutazione della bontà della stima, che è spesso basata sulla minimizzazione di una misura di errore che si commette utilizzando $\hat{f}$ quando la densità reale sottostante i dati è f. Per la valutazione della bontà della stima puntuale, la misura di divergenza più comunemente utilizzata è l'errore quadratico medio (*Mean Squared Error*, MSE), definito come la media dei quadrati degli errori:

$$MSE_{\boldsymbol{x}}(\hat{f}_h(\boldsymbol{x})) = E\left[(\hat{f}_h(\boldsymbol{x}) - f(\boldsymbol{x}))^2\right].$$

Nel caso in cui si desiderasse valutare la bontà dello stimatore globalmente, gli approcci più utilizzati sono basati sull'errore quadratico integrato (*Integrated Squared Error*, ISE), definito come l'integrazione su $\mathbb{R}^d$ degli errori al quadrato:

$$ISE(\hat{f}_h) = \int_{\mathbb{R}^d} (\hat{f}_h(\boldsymbol{x}) - f(\boldsymbol{x}))^2 d\boldsymbol{x}.$$

L'ISE dipende dal campione, quindi è una quantità casuale. Per questo motivo si ricorre usualmente al suo valore atteso, l'errore quadratico medio integrato (*Mean Integrated Squared Error*, MISE), definito da:

$$MISE(\hat{f}_h) = E\left[\int_{\mathbb{R}^d} (\hat{f}_h(\boldsymbol{x}) - f(\boldsymbol{x}))^2 d\boldsymbol{x}\right]. \tag{1.4}$$

Poiché la quantità da integrare è non negativa per ogni $\boldsymbol{x}$, l'ordine tra valore atteso e integrale può essere invertito. Così facendo, dopo alcuni sviluppi, la (1.4) può essere scritta come segue:

$$\begin{aligned} MISE(\hat{f}_h) &= \int_{\mathbb{R}^d} MSE(\hat{f}_h(\boldsymbol{x}))d\boldsymbol{x} = \int_{\mathbb{R}^d} E\{\hat{f}_h(\boldsymbol{x}) - f(\boldsymbol{x})\}^2 d\boldsymbol{x} \\ &= \int_{\mathbb{R}^d} \{E[\hat{f}_h(\boldsymbol{x})] - f(\boldsymbol{x})\}^2 d\boldsymbol{x} + \int_{\mathbb{R}^d} Var[\hat{f}_h(\boldsymbol{x})]d\boldsymbol{x} \\ &= \int_{\mathbb{R}^d} bias^2(\hat{f}_h(\boldsymbol{x}))d\boldsymbol{x} + \int_{\mathbb{R}^d} Var[\hat{f}_h(\boldsymbol{x})]d\boldsymbol{x}. \end{aligned}$$

Per questioni di semplicità matematica, viene spesso considerata la versione asintotica del MISE (*Asymptotic Mean Integrated Squared Error*, AMISE), che risulta anche essa dalla somma di una componente legata alla varianza e una legata alla distorsione. Si assuma che il *kernel* K sia una funzione di densità radialmente simmetrica e che ammetta derivate parziali di secondo ordine continue e limitate. Definite

$$\mu_2(K)\boldsymbol{I}_d = \int_{\mathbb{R}^d} \boldsymbol{t}\boldsymbol{t}^\top K(\boldsymbol{t})d\boldsymbol{t} \qquad e \qquad R(K) = \int_{\mathbb{R}^d} K(\boldsymbol{t})^2 d\boldsymbol{t},$$

tramite uno sviluppo di Taylor del secondo ordine si ottiene:

$$\begin{aligned} bias[\hat{f}_h(\boldsymbol{x})] &\approx \frac{1}{2}h^2\mu_2(K)\nabla^2 f(\boldsymbol{x}) \\ Var[\hat{f}_h(\boldsymbol{x})] &\approx n^{-1}h^{-d}R(K)f(\boldsymbol{x}). \end{aligned}$$

Dunque, l'AMISE, nel caso di una densità d-variata, risulta essere:

$$AMISE(\hat{f}) = \frac{h^4}{4}\mu_2(K)^2 \int_{\mathbb{R}^d} \{\nabla^2 f(\boldsymbol{x})\}^2 d\boldsymbol{x} + \frac{1}{nh^d}R(K). \qquad (1.5)$$

Pertanto, la scelta del parametro di lisciamento rientra nel contesto più generale del compromesso distorsione-varianza, in quanto valori piccoli sono associati a varianze maggiori e bias minori, mentre valori grandi sono associati a varianze minori e bias maggiori. Si veda, per un'illustrazione, la Figura 1.3.

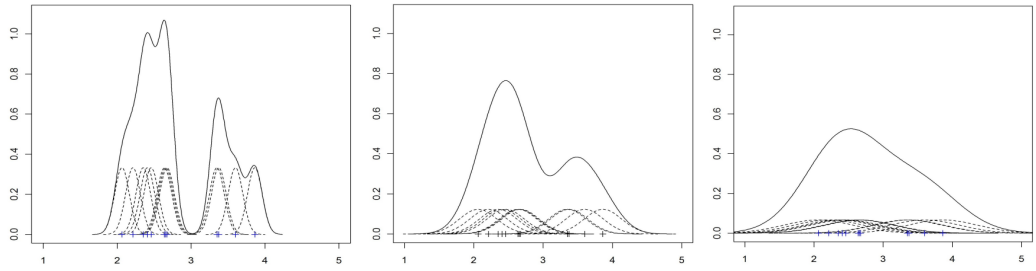

Figura 1.3: Densità stimata al variare di h. A sinistra la curva è sovra-adattata, al centro il lisciamento è eccessivo, mentre a destra è ottimale.

Un possibile criterio per la selezione del parametro di lisciamento consiste nel selezionare h in modo tale da garantire la minimizzazione del MISE. Considerando l'AMISE, invece, si ottiene la seguente espressione:

$$h_{AMISE} = n^{-\frac{1}{5}} \left\{ \frac{\int K^2(\boldsymbol{x}) d\boldsymbol{x}}{\int [f''(\boldsymbol{y})]^2 [\int \boldsymbol{x}^2 K(\boldsymbol{x}) d\boldsymbol{x}]^2 d\boldsymbol{y}} \right\}. \tag{1.6}$$

In alcune circostanze h_{AMISE} rappresenta una buona approssimazione del parametro di lisciamento ottenuto minimizzando il MISE (Marron e Wand, 1992).

La definizione del parametro di lisciamento dipenderà sempre dalla vera funzione di densità, per cui sono stati proposti metodi *data driven* o metodi che si basano su una famiglia parametrica, volti a sostituire $f(\boldsymbol{x})$ con una sua stima pilota. Fra questi ultimi, la regola del pollice risale a Deheuvels (1977), ma fu resa nota grazie a Silverman (1986). L'idea consiste nel sostituire la parte ignota di h_{AMISE} con una stima basata su un modello parametrico noto, come quello gaussiano, e nello scegliere il parametro asintoticamente ottimale sotto il modello assunto. Tuttavia, il parametro di lisciamento selezionato risulterà appropriato solo nel caso in cui la densità presa come riferimento si avvicina alla vera densità f. Nel contesto delle tecniche più prettamente *data driven*, invece, si menzionano metodi *bootstrap* o di

convalida incrociata. Per una trattazione dettagliata si vedano Wand e Jones (1994), Bowman e Azzalini (1999).

Stimatori del nucleo adattivi

Lo stimatore fin qui discusso utilizza un parametro di lisciamento fisso, globale. Tuttavia, in alcune circostanze è preferibile utilizzare un parametro di lisciamento variabile, in modo tale da ottenere un risultato ragionevole anche quando vi siano regioni con una maggiore concentrazione rispetto ad altre, come nel contesto in esame. In particolare, è opportuno lisciare meno dove la densità è più elevata e lisciare maggiormente in regioni a densità inferiore. Al fine di far variare l'ampiezza di banda, sono state proposte due alternative (Terrell e Scott, 1992). La prima consiste nel far variare il parametro di lisciamento con il punto $\boldsymbol{x}$ nel quale si vuole stimare la densità e non con l'osservazione: si parla in questo caso di *baloon estimator*.

$$\hat{f}_B(\boldsymbol{x}) = \frac{1}{nh(\boldsymbol{x})^d} \sum_{i=1}^{n} K\left(\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h(\boldsymbol{x})}\right). \tag{1.7}$$

Una seconda proposta è il *sample point estimator* (Breiman *et al.*, 1977), che consiste nel far variare il parametro di lisciamento h con l'osservazione i-esima. Si tratta quindi di una mistura di funzioni *kernel* centrate in ogni osservazione e riscalate individualmente:

$$\hat{f}_h(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h(\boldsymbol{x}_i)^d} K\left(\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h(\boldsymbol{x}_i)}\right). \tag{1.8}$$

Il primo stimatore è computazionalmente più efficiente e, fissato $\boldsymbol{x}$, ha il vantaggio di mantenere invariate le proprità dello stimatore *kernel* tradizionale, ma non integra a uno. Il secondo, invece, è una densità, ma la stima risulta essere influenzata anche da osservazioni lontane dal punto in cui vuole essere valutata la densità.

Tra le proposte per selezionare $h(\boldsymbol{x}_i)$ in uno stimatore *sample point* Abramson (1982) suggerisce

$$h_i^A = h(\boldsymbol{x}_i) = h_0 \left[\tilde{f}_{h_0}(\boldsymbol{x}_i)\right]^{-1/2}, \tag{1.9}$$

mentre la regola di selezione introdotta da Silverman (1986) è data da

$$h_i^S = h_0 \left[\frac{1}{g}\tilde{f}_{h_0}(\boldsymbol{x}_i)\right]^{\beta}, \tag{1.10}$$

dove $\tilde{f}_{h_0}$ è una densità pilota stimata con un parametro di lisciamento h_0 scalare, g è la media geometrica di $f_{h_0}(\boldsymbol{x}_i)$, con i=1,...,n e $\beta \in [0,1]$ è un parametro che regola la sensibilità dell'ampiezza di banda in base alla densità pilota.

1.2.3 Stima non pametrica *binned*

Quando i dati in esame sono raggruppati in classi, o al fine di ridurre il numero di operazioni di calcolo in presenza di campioni di dimensioni considerevoli, è utile adottare un approccio basato sulla stima *kernel binned*. Nel caso unidimensionale, una regola di *binning* può essere rappresentata da una sequenza di funzioni $\{w_j(x,\delta), j \in \mathbb{Z}\}$ e richiede che le osservazioni siano distribuite tra i punti di una griglia $g_j = j\delta$ in cui ad ogni cella g_j è assegnato il peso $w_j(x,\delta)$. Se richiediamo che per ogni x e $\delta > 0$, $\sum_j w_j(x,\delta) = 1$, una regola di *binning* divide i dati assegnandoli a differenti punti della griglia. Un esempio di regola di *binning* è il *binning* semplice, dove

$$w_j(x,\delta) = \begin{cases} 1 & \text{se} \quad x \in \left(\left(j-\frac{1}{2}\right)\delta, \left(j+\frac{1}{2}\right)\delta\right] \\ 0 & \text{altrimenti.} \end{cases} \tag{1.11}$$

Dunque, si costruisce una griglia equispaziata, con intervalli di ampiezza δ. Dato t_j il centro del j-esimo *bin*, il quale contiene n_j osservazioni, si ha $t_{j+1} - t_j = \delta$ e $\sum_{j\in\mathbb{Z}} n_j = n$, con n numero di os-

servazioni. Lo stimatore *kernel binned* è così ottenuto:

$$\tilde{f}_h(x) = \frac{1}{nh} \sum_{j \in \mathbb{Z}} n_j K\left(\frac{x - t_j}{h}\right). \tag{1.12}$$

Regole di *binning* multivariate possono essere definite prendendo il prodotto di regole univariate, come segue. Si supponga che $w_{ij}(x, \delta)$ denoti una regola univariata per ogni $1 \leq i \leq d$. Siano $j = (j_1, ..., j_d)$, $\boldsymbol{x} = (x_1, ...x_d)$ e $\boldsymbol{\delta} = (\delta_1, ...\delta_d)$ vettori d-dimensionali e definiamo il punto della griglia $\boldsymbol{t}_j$ con $\boldsymbol{t}_j = j\boldsymbol{\delta} \equiv (j_1\delta_1, ..., j_d\delta_d)$. La regola di *binning* diventa:

$$w_j^d(\boldsymbol{x}, \boldsymbol{\delta}) = \prod_{i=1}^{d} w_{ij_i}(x_i, \delta_i). \tag{1.13}$$

Allo stesso modo, stimatori *kernel* d-variati possono essere definiti moltiplicativamente. La versione *binned* dello stimatore in ambito multidimensionale è data da:

$$\tilde{f}_h(\boldsymbol{x}) = \frac{1}{n} \sum_{j \in \mathbb{Z}^d} n_j K_h^d\left(\boldsymbol{x} - \boldsymbol{t}_j\right). \tag{1.14}$$

Una questione di rilevanza pratica riguarda l'accuratezza dello stimatore *kernel* basato su dati suddivisi in *bin*. Siano $\tilde{f}$ lo stimatore *kernel binned* per la densità f e $\hat{f}$ lo stimatore *kernel* ordinario. L'accuratezza di $\tilde{f}$ può essere valutata in due modi differenti. Il primo consiste nel trattare $\tilde{f}$ come uno stimatore di f e studiarne le proprietà come si fa usualmente con gli stimatori *kernel* ordinari. Il secondo prende in considerazione la vicinanza tra $\hat{f}$ e $\tilde{f}$. Questi due obiettivi possono essere differenti, in quanto una regola di *binning* che porta a migliori proprietà di stima di $\tilde{f}$ non necessariamente conduce ad una vicinanza maggiore di $\tilde{f}$ a $\hat{f}$. Il secondo approccio è appropriato, in quanto $\hat{f}$ è lo stimatore più naturale e matematicamente semplice dei due, mentre, per ragioni di efficienza computazionale, $\tilde{f}$ è lo stimatore più conveniente in pratica. Inoltre, gran parte del contributo teorico

si riferisce a $\hat{f}$ piuttosto che $\tilde{f}$, quindi è di particolare interesse capire quanto le stime siano vicine l'una con l'altra. In generale, la scelta della dimensione della griglia è di primaria importanza, dal momento che è legata al compromesso tra la minimizzazione dell'errore dovuto all'operazione di *binning* e la minimizzazione del tempo impiegato per la stima. Una misura che permette di tenere conto dell'errore dovuto al *binning* è il *Relative Mean Integrated Squared Error*, RMISE, definito come:

$$RMISE = \frac{E\left[\int_{\mathbb{R}^d}(\tilde{f}_{h_0}(\boldsymbol{x}) - \hat{f}_{h_0}(\boldsymbol{x}))^2 d\boldsymbol{x}\right]}{E\left[\int_{\mathbb{R}^d}(\hat{f}_{h_0}(\boldsymbol{x}) - f(\boldsymbol{x}))^2 d\boldsymbol{x}\right]},$$

dove h_0 è il parametro di lisciamento che minimizza $E[\int_{\mathbb{R}^d}(\hat{f}(\boldsymbol{x}) - f(\boldsymbol{x}))^2 d\boldsymbol{x}]$. Minimizzando la quantità introdotta, l'utilizzo della stima *kernel binned* ha un effetto minimo sull'errore del processo di stima.

Per approfondimenti sulla stima non parametrica *binned* e sulla sua accuratezza, si vedano Scott e Sheather (1985) e Hall e Wand (1994).

1.3 Individuazione dei gruppi

Il *clustering* non parametrico si fonda sulla premessa che i gruppi corrispondano ai domini di attrazione delle mode della densità f sottostante i dati (Stuetzle, 2003). Dopo aver opportunamente stimato la densità f, l'obiettivo diventa trovare le mode e assegnare ogni osservazione al dominio di attrazione di una moda. A questo scopo, il *clustering* non parametrico si è sviluppato in due principali direzioni, le quali presentano delle differenze operative nell'individuazione delle mode (Menardi, 2016).

Una prima direzione si basa sull'individuazione esplicita delle mode della densità e, conseguentemente, sull'associazione di ogni osservazione alla moda di pertinenza. Gran parte dei contributi sviluppati in questa direzione si basa su un meccanismo di risalita del gradiente, spesso basato sull'algoritmo *mean-shift* (Fukunaga e Hostetler, 1975),

che consiste nel muovere ciascuna osservazione lungo il percorso più ripido di risalita della funzione di densità, fino alla convergenza ad un ottimo locale, ovvero una moda. Ogni *cluster* è dunque definito come il luogo dei punti in cui il gradiente punta ad una stessa moda di f.

In alternativa, una via indiretta, introdotta da Hartigan (1975) e che non consiste nell'associazione esplicita dei *cluster* alle mode della distribuzione, identifica i gruppi tramite regioni dello spazio campionario ad elevata densità, definite dalle sue curve di livello. Formalmente, la regione dello spazio la cui densità è superire ad un fissato valore λ è definita da

$$L(\lambda; f) = \left\{ \boldsymbol{x} \in \mathbb{R}^d : f(\boldsymbol{x}) \geq \lambda \right\}, 0 \leq \lambda \leq \max(f), \tag{1.15}$$

e può essere connessa o disconnessa a seconda del valore fissato per *lambda*. Nel caso in cui la regione sia sconnessa, essa sarà formata da un numero di componenti connesse, ognuna delle quali associata ad un gruppo al livello λ. L'idea su cui si fonda questo approccio è che ogni regione connessa di $L(\lambda)$ contiene almeno una moda della densità e che, per ogni moda, esiste un valore di λ per cui una delle componenti connesse del relativo insieme di livello contiene solo quella moda. Poiché nella pratica $f(\boldsymbol{x})$ è ignota, una stima $\hat{L}(\lambda)$ di $L(\lambda)$ è ottenuta sostituendo nella (1.15) $f(\boldsymbol{x})$ con una sua stima non parametrica $\hat{f}(\boldsymbol{x})$.

In generale, non è garantito che esista un singolo valore di $\lambda = \lambda^*$, in grado di individuare tutte le mode della densità, cioè tale per cui ogni moda appartenga ad una distinta componente connessa di $L(\lambda)$. Ciononostante, l'intera struttura modale può essere ricostruita facendo variare λ nell'intervallo di valori che esso può assumere, andando in questo modo a determinare una struttura gerarchica, l'albero dei *cluster*, che conta il numero di componenti connesse degli insiemi di livello al variare di λ, come si vede in Figura 1.4. Ciascuna delle sue foglie corrisponde ad un *cluster*, cioè la più grande componente connessa di $L(\lambda)$ che include una singola moda. Il nodo radice rappresenta l'intero supporto ed è associato a $\lambda = 0$, mentre per determinare

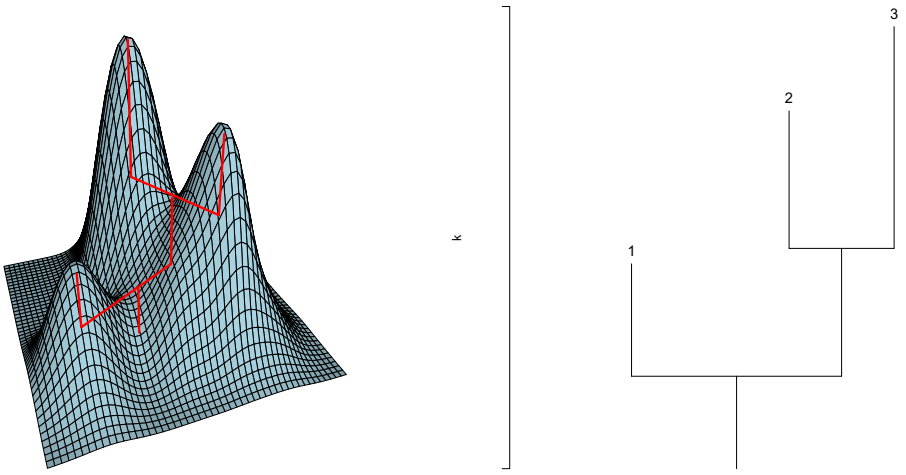

Figura 1.4: Un esempio di una distribuzione bivariata trimodale e relativo albero dei *cluster*.

i figli del nodo radice è necessario individuare il valore minimo di λ per cui la curva di livello risulta essere disconnessa. Nel caso in cui non esistesse un λ che soddisfi questa condizione, la densità sarebbe unimodale.

Un'illustrazione di tali concetti è fornita nella Figura 1.5 per un insieme di dati bidimensionali tratti da una funzione di densità trimodale. Al livello λ_1 l'insieme di livello si divide in due componenti, una delle quali associata univocamente ad una delle mode e rappresenta un primo nucleo di un *cluster*. Le mode più alte sono contenute nella seconda componente. Al livello λ_2, corrispondente alla densità della moda più bassa, la prima componente scompare, mentre la seconda rimane e si divide ulteriormente ad un livello λ_3 in due componenti, ciascuna associata ad una singola moda.

Uno dei fattori che hanno limitato l'applicazione di questo metodo consiste nel fatto che l'individuazione delle componenti connesse di $L(\lambda)$ rimane concettualmente e computazionalmente semplice solamente nel caso unidimensionale, dove gli insiemi connessi sono intervalli. In spazi multidimensionali, il problema è stato usualmente

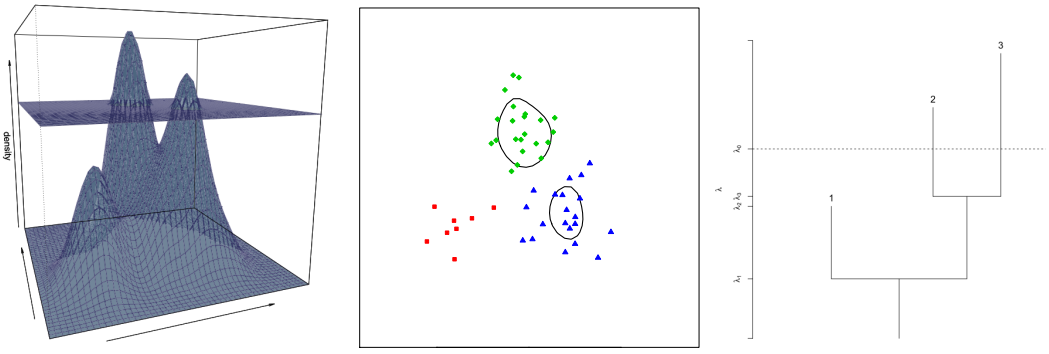

Figura 1.5: Una sezione della funzione di densità in Figura 1.4 ad un livello λ_0, l'insieme di livello identificato (centro), formato da due regioni disconnesse. A sinistra l'albero dei *cluster* in cui si evidenzia il livello λ_0.

affrontato tramite il ricorso alla teoria dei grafi. Infatti, poiché l'interesse primario è attribuire le osservazioni ai *cluster*, piuttosto che identificare una partizione dell'intero spazio $\mathbb{R}^d$, la costruzione di un opportuno grafo $\mathscr{G}$, i cui vertici sono $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$, e la successiva identificazione delle componenti connesse di $\mathscr{G}$ permette di semplificare il problema in esame riconducendosi da uno spazio multidimensionale continuo. Qualora il grafo non sia connesso, ogni sottografo connesso identifica una componente connessa. Dunque, il sottografo $\mathscr{G}_{\lambda^*}$ indotto dall'insieme di livello campionario

$$\mathscr{S}(\lambda^*) = \{\boldsymbol{x}_i \in (\boldsymbol{x}_1, ..., \boldsymbol{x}_n) : \hat{f}(\boldsymbol{x}_i) \geq \lambda^*\} \tag{1.16}$$

viene costruito eliminando da $\mathscr{G}$ tutti i vertici non contenuti in $\mathscr{S}(\lambda^*)$ e tutti gli archi con almeno un vertice tra di essi. Questa operazione dovrebbe essere ripetuta $\forall \lambda : 0 \leq \lambda \leq \max(\hat{f})$, ma in pratica viene considerata una griglia finita di valori per λ. Le componenti connesse di $\mathscr{G}_{\lambda^*}$ costituiscono un'approssimazione di quelle di $L(\lambda^*)$, per cui è possibile costruire l'albero dei *cluster*. Il problema viene così semplificato riconducendosi da uno spazio multidimensionale continuo ad uno spazio finito e discreto.

In letteratura sono state avanzate proposte differenti per la costruzione del grafo $\mathscr{G}$. Azzalini e Torelli (2007) utilizzano la *triangola-*

zione di Delaunay, il grafo associato alla *tassellatura di Voronoi*, che suddivide lo spazio $\mathbb{R}^d$ in n regioni, contenenti esattamente un'osservazione e i cui punti sono più prossimi all'osservazione in quella determinata regione rispetto a qualunque altra. La triangolazione di Delaunay viene costruita collegando tramite un arco le osservazioni le cui regioni definite dalla tassellatura di Voronoi condividono una faccia del poliedro. Al fine di superare il problema legato alla complessità computazionale del metodo descritto, che cresce esponenzialmente con d, Menardi e Azzalini (2014) propongono un metodo che sfrutta caratteristiche della densità per la costruzione di $\mathscr{G}$. In particolare, viene costruito un arco tra vertici quando $\hat{f}$ non presenta avvallamenti lungo il segmento che li unisce.

La procedura fin qui descritta fornisce M gruppi di punti, chiamati nuclei dei *cluster* e lascia le osservazioni rimanenti, sulle code della distribuzione o in corrispondenza della valle tra due mode, prive di un'etichetta di appartenenza ad un gruppo. Azzalini e Torelli (2007) utilizzano un metodo di classificazione per allocare queste osservazioni: per ogni generica osservazione non allocata, $\boldsymbol{x}_0$, si calcola la stima di densità $\hat{f}_j(\boldsymbol{x}_0)$ basata sulle osservazioni precedentemente assegnate al gruppo j, $j = 1,...,M$ e si assegna $\boldsymbol{x}_0$ al gruppo per cui la quantità $\hat{f}_j(\boldsymbol{x}_0)/\max_{k\neq j}\hat{f}_k(\boldsymbol{x}_0)$ è massima.

Capitolo 2

Clustering non parametrico per dati direzionali

2.1 Dati direzionali

Come anticipato, il dato di cui si dispone, il fotone, ha natura direzionale. I dati direzionali sono osservazioni definite come vettori su un'ipersfera di dimensione q, Ω_q. Idealmente, il LAT è al centro del sistema di coordinate e l'emissione dei fotoni avviene intorno ad esso, come illustrato in Figura 2.1, mentre i dati a disposizione rappresentano proiezioni su una sfera, per cui hanno natura direzionale. Nel contesto della rilevazione di sorgenti di raggi gamma, i dati tipicamente consistono nella direzione nello spazio di ogni fotone rilevato, insieme ad informazioni aggiuntive. Le direzioni nello spazio tridimensionale possono essere rappresentate tramite le coordinate cartesiane come vettori $\boldsymbol{x}$ di norma unitaria, cioè punti sulla sfera

$$\Omega_2 = \{\boldsymbol{x} \in \mathbb{R}^3 : \|\boldsymbol{x}\|_2 = x_1^2 + x_2^2 + x_3^2 = 1\} \subset \mathbb{R}^3.$$

Le coordinate cartesiane possono essere ottenute a partire dalle coordinate galattiche, ovvero a partire dalla longitudine $l \in (-180, +180)$ e dalla latitudine $b \in (-90, +90)$, con

$$\boldsymbol{x} = [\cos(l)\cos(b), \sin(l)\cos(b), \sin(b)]^\top.$$

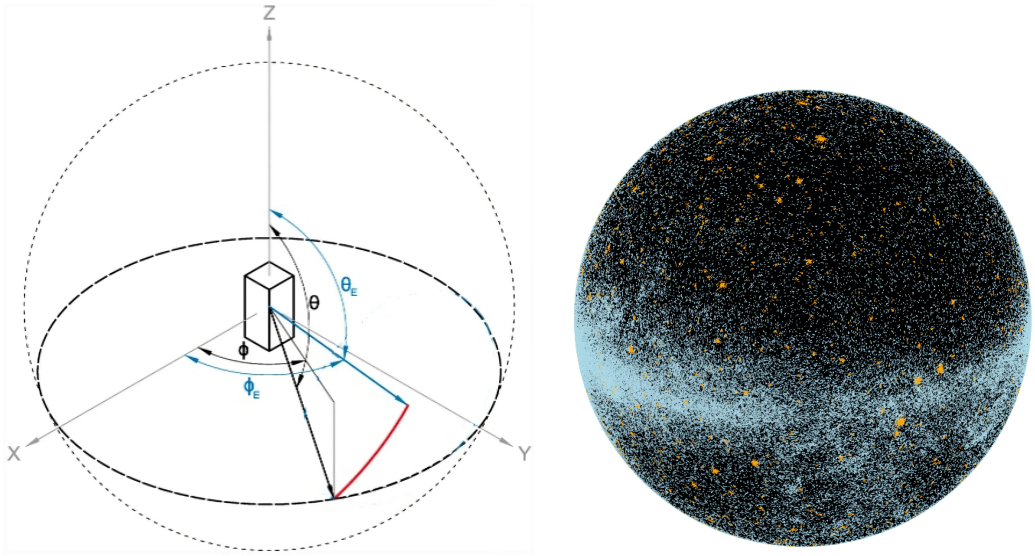

Figura 2.1: A sinistra, rappresentazione del sistema di coordinate. A destra, rappresentazione sferica del *background* (azzurro) e delle sorgenti (arancione).

É opportuno tenere in considerazione che fotoni rilevati ai bordi opposti della mappa espressa tramite latitudine e longitudine sono in realtà spazialmente vicini. Ad esempio, il punto (0°, 0°) coincide con il punto (360°, 0°). Per questo motivo, si sceglie di lavorare sul sistema di coordinate cartesiane.

In questo caso le osservazioni a disposizione, in seguito indicate con $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$, possono essere considerate facenti parte di un campione derivante da una funzione di densità che ha come supporto la sfera. Inoltre, le mode della densità rimangono associabili a *cluster*, dunque la tecnica di individuazione dei gruppi introdotta nel capitolo precedente può essere naturalmente estesa. La natura direzionale dei dati spinge a riadattare al contesto in esame il processo di stima della densità, ma, al contempo, è possibile sfruttare la loro contenuta dimensionalità per proporre soluzioni computazionalmente efficienti.

2.2 La griglia

Per una più efficiente gestione della mole di dati a disposizione e dello sforzo computazionale che ne consegue, negli sviluppi che seguono si ricorrerà ad una opportuna riduzione del numero di operazioni richieste attraverso la suddivisione dei dati in una griglia adeguata. Come sarà chiarito nelle sezioni che seguono, tale accorgimento sarà utile sia a rendere più efficiente la determinazione della stima *kernel*, sia la successiva individuazione dei gruppi.

La letteratura propone diverse procedure per costruire una griglia su una sfera (Wenninger, 1979; Williams, 1979). Una prima via consiste nel tagliare la sfera lungo i meridiani e i paralleli. I primi sono dati dall'intersezione tra la superficie sferica e piani passati per i due poli, mentre i secondi sono circonferenze immaginarie ottenute dall'intersezione tra la superficie sferica e piani perpendicolari alla linea passante per i due poli. I vertici delle celle saranno dati dall'intersezione tra meridiani e paralleli. Un secondo modo consiste nel costruire una griglia poligonale, ovvero le cui celle sono poligoni, come triangoli o quadrilateri. In questo contesto si utilizza il secondo approccio ed, in particolare, viene costruita una griglia triangolare, in quanto la distribuzione dei vertici è molto più regolare. Se utilizzassimo meridiani e paralleli per costruire la griglia, infatti, otterremmo delle celle non uniformi, di dimensione inferiore ai poli. Nel contesto in esame, sarebbe invece ideale ottenere un risultato opposto: le celle dovrebbero avere dimensione inferiore nella regione lungo il piano galattico, caratterizzato dal un'elevata presenza di osservazioni provenienti dal *background* diffuso ed in cui sarebbe opportuno poter distinguere sorgenti molto vicine, evitando che giacciano all'interno di una stessa cella. La griglia, dunque, viene costruita a partire da un icosaedro, i cui spigoli vengono ricorsivamente divisi, in modo tale che le facce triangolari vengano suddivise in più triangoli. Infine si considerano le proiezioni sulla superficie sferica. In questo modo i

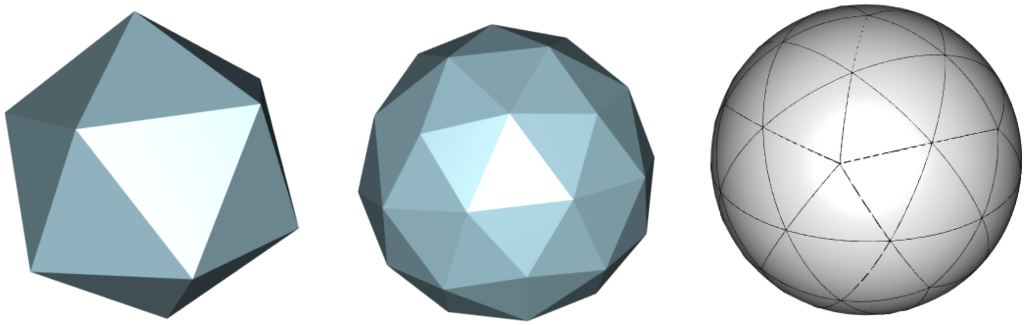

Figura 2.2: Rappresentazione di un icosaedro (sinistra), la figura ottenuta suddividendo in due ogni suo spigolo (centro) e la proiezione sulla sfera (destra).

centroidi di ogni cella giaceranno sulla sfera stessa e non sarà necessario approssimare i rappresentanti di ogni *bin* con punti della griglia non appartenenti alla sfera. Per un'illustrazione si veda la Figura 2.2.

2.3 Stima non parametrica per dati direzionali

2.3.1 Definizione del nucleo direzionale

Siano $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ osservazioni indipendenti ed identicamente distribuite da una funzione di densità f su Ω_2 tale che

$$\int_{\Omega_2} f(\boldsymbol{x})\omega_2(d\boldsymbol{x}) = 1,$$

dove ω_2 denota la misura di Lebesgue su Ω_2. In particolare, l'area di Ω_q è data da

$$\omega_q = \omega_q(\Omega_q) = \frac{2\pi^{\frac{q+1}{2}}}{\Gamma(\frac{q+1}{2})}, \quad q \geq 1,$$

dove $\Gamma(p) = \int_0^\infty x^{p-1}e^{-x}dx$.

Lo stimatore *kernel* per dati direzionali è stato proposto da Hall *et al.* (1987) e Bai *et al.* (1988), seguendo due prospettive differenti.

Si farà riferimento al secondo, dato da

$$\hat{f}_h(x) = \frac{c_h(K)}{n} \sum_{i=1}^{n} K\left(\frac{1 - \boldsymbol{x}^\top \boldsymbol{x}_i}{h^2}\right) \tag{2.1}$$

con $K(\cdot)$ nucleo per dati direzionali, decrescente in $[0,\infty)$ e $c_{h,q}(K)$ definita da

$$c_h(K)^{-1} = \int_{\Omega_2} K\left(\frac{1 - \boldsymbol{x}^\top \boldsymbol{x}_i}{h^2}\right) \omega_2(d\boldsymbol{x}) = h^2 \lambda_{h,2}(K),$$

dove $\lambda_{h,q}(K) = \omega_{q-1} \int_0^{2h^{-2}} K(r) r^{\frac{q}{2}-1} (2 - rh^2)^{\frac{q}{2}-1} dr$. Una distribuzione largamente diffusa per modellare l'emissione di raggi gamma nel campo astrofisico è la distribuzione di von Mises-Fisher (vMF) (Mardia e Jupp, 2000)

$$f_{vMF}(\boldsymbol{x}; \boldsymbol{\mu}, \kappa) = C_2(\kappa) e^{\kappa \boldsymbol{x}^\top \boldsymbol{\mu}}, \tag{2.2}$$

che estende al caso sferico la distribuzione normale tridimensionale $N_3(\boldsymbol{\mu}, \kappa^{-1} \boldsymbol{I}_3)$, con $\boldsymbol{I}_3$ matrice identità 3×3. Il parametro $\boldsymbol{\mu}$ è direttamente legato alla direzione media e ha norma unitaria, mentre $\kappa \geq 0$ è un parametro di concentrazione che indica quanto l'emissione di fotoni si distribuisce attorno alla direzione media della sorgente. Un caso particolare di questa densità si ha per $\kappa = 0$, per cui la distribuzione vMF si riduce ad una distribuzione uniforme sulla sfera di raggio unitario, che assegna probabilità ω_q^{-1} ad ogni direzione di Ω_q, mentre collassa ad un punto di massa centrato in $\boldsymbol{\mu}$ se $\kappa \to \infty$. La costante di normalizzazione è data da

$$C_2(\kappa) = \frac{\kappa^{\frac{1}{2}}}{(2\pi)^{\frac{3}{2}} \mathscr{I}_{\frac{1}{2}}(\kappa)}, \tag{2.3}$$

contenente la funzione Bessel modificata

$$\mathscr{I}_\nu(z) = \frac{\left(\frac{z}{2}\right)^\nu}{\pi^{1/2} \Gamma(\nu + \frac{1}{2})} \int_{-1}^{1} (1 - t^2)^{\nu - \frac{1}{2}} e^{zt} dt \tag{2.4}$$

di ordine $\nu = 1/2$. Utilizzando il *kernel* von Mises-Fisher la (2.1) diventa

$$\begin{aligned}\hat{f}_h(\boldsymbol{x}) &= \frac{1}{n}\sum_{i=1}^{n} f_{vMF}\left(\boldsymbol{x}; \boldsymbol{x}_i, \frac{1}{h^2}\right) \\ &= \frac{1}{(2\pi)^{\frac{3}{2}}\mathscr{I}_{\frac{1}{2}}(h^{-2})}\frac{1}{hn}\sum_{i=1}^{n}\exp\left(\frac{\boldsymbol{x}^\top \boldsymbol{x}_i}{h^2}\right),\end{aligned} \tag{2.5}$$

cioè lo stimatore *kernel* per dati direzionali sulla sfera di raggio unitario è una mistura di distribuzioni von Mises-Fisher tridimensionali con $\kappa = \frac{1}{h^2}$.

2.3.2 Il parametro di lisciamento

Criteri di selezione

La scelta di un parametro di lisciamento adeguato rimane una questione sostanziale per poter garantire un'affidabile separazione delle sorgenti dal *background*.

Per poter procedere, è necessario fornire una riscrittura della formulazione (2.5), in quanto il parametro di lisciamento minimo che consente la determinazione numerica della stima, trovato attraverso il software R (R Core Team, 2022), non restituisce una stima adeguatamente precisa della densità, dunque ci servirà un valore di h più prossimo allo zero. I problemi computazionali sorgono in quanto per $h \to 0^+$ si ha $\kappa \to \infty$, la funzione di Bessel è tale per cui $\mathscr{I}_\nu(\cdot) \to \infty$ e la costante di normalizzazione $C_2(\kappa) \to 0$. Inoltre, $\kappa \boldsymbol{x}^\top \boldsymbol{\mu}$, per $\kappa \to \infty$, va a infinito e cresce ulteriormente a causa dell'applicazione dell'esponenziale. Dopo alcuni sviluppi e trasformazioni (Montin, 2021) si ottiene la seguente riscrittura della (2.5) che permette l'utilizzo di un

parametro di lisciamento prossimo a zero:

$$\begin{aligned} f(y) &= \exp\left(\log f(y)\right) \\ &= \exp\left[a + \log \sum_i \exp\left(\frac{y_i}{h^2} - a\right)\right. \\ &\qquad \left. - 2\log(h) - \log(2\pi) - \frac{1}{h^2}\right], \end{aligned} \tag{2.6}$$

dove $y_i = \boldsymbol{x}^\top \boldsymbol{x}_i$ e $a = max\left(\frac{y_i}{h^2}\right)$.

Il problema della scelta di un parametro opportuno per dati direzionali è stato affrontato soprattutto nel caso circolare, ovvero nel caso in cui i dati appartengano allo spazio Ω_1. In generale, i metodi di selezione introdotti nel capitolo precedente ammettono un'estensione per dati direzionali. Il MISE per lo stimatore *kernel* direzionale è dato da

$$MISE(\hat{f}) = E\left[ISE(\hat{f})\right] = E\left[\int_{\Omega_2} \left(\hat{f}_h(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2 \omega_2(d\boldsymbol{x})\right]$$

e anche in questo caso è possibile scegliere il parametro di lisciamento in modo tale da garantire la minimizzazione del MISE:

$$h_{MISE} = \underset{h>0}{\arg\min} MISE(\hat{f}).$$

Le prime proposte per la selezione del parametro di lisciamento basate su un approccio di tipo *data driven* furono avanzate da Hall *et al.* (1987), mentre in Taylor (2008) è possibile trovare un metodo di selezione basato sulla regola del *plug-in* nel caso circolare. Per la scelta della *bandwidth* adatta, si può partire dalla definizione di h_{AMISE} in (1.6). García-Portugués (2013) propone la seguente regola

del pollice per dati direzionali:

$$h_{ROT} = \begin{cases} \left[\dfrac{4\pi^{\frac{1}{2}}\mathscr{I}_0(\hat{\kappa})^2}{\hat{\kappa}[2\mathscr{I}_1(2\hat{\kappa})+3\hat{\kappa}\mathscr{I}_2(2\hat{\kappa})]n}\right]^{\frac{1}{5}}, & q=1, \\ \left[\dfrac{8\sinh^2(\hat{\kappa})}{\hat{\kappa}[(1+4\hat{\kappa}^2)\sinh(2\hat{\kappa})-2\hat{\kappa}\cosh(2\hat{\kappa})]n}\right]^{\frac{1}{6}}, & q=2, \\ \left[\dfrac{4\pi^{\frac{1}{2}}\mathscr{I}_{\frac{q-1}{2}}(\hat{\kappa})^2}{\hat{\kappa}^{\frac{q+1}{2}}\left[2q\mathscr{I}_{\frac{q+1}{2}}(2\hat{\kappa})+(2+q)\hat{\kappa}\mathscr{I}_{\frac{q+3}{2}}(2\hat{\kappa})\right]n}\right]^{\frac{1}{4+q}}, & q\geq 3. \end{cases} \quad (2.7)$$

dove il parametro di concentrazione $\hat{\kappa}$ è stimato tramite massima verosimiglianza. Tuttavia, in questo contesto non lavoreremo simultaneamente su tutta la sfera, bensì su regioni indipendenti separate da regioni vuote. Poiché, nella maggior parte dei casi, le zone trovate corrispondono a porzioni limitate di sfera, la stima di massima verosimiglianza del parametro di concentrazione κ tenderebbe ad infinito e crescerebbe ulteriormente a seguito dell'applicazione dell'esponenziale contenuto nelle funzioni iperboliche. Considerando quanto detto e ricordando che per $\kappa \to \infty$ la distribuzione von Mises-Fisher è approssimabile da una distribuzione normale, si è scelto di utilizzare un parametro di lisciamento su scala normale, spesso utilizzato per la sua semplicità in termini di complessità matematica e computazionale (Silverman, 1986). L'ignota densità f è sostituita da una densità normale di media $\boldsymbol{\mu}$ e varianza Σ. Seguendo quanto proposto da Chacón e Duong (2018), si può dimostrare che l'*AMISE* è minimizzato prendendo

$$h_{NS} = \left[\frac{4d|\Sigma|^{\frac{1}{2}}}{2\mathrm{tr}(\Sigma^{-2})+\mathrm{tr}^2(\Sigma^{-1})}\right]^{\frac{1}{d+4}} n^{-\frac{1}{d+4}}, \quad (2.8)$$

dove Σ è sostituita con una sua stima non distorta. Dato che la densità normale è una fra le densità più lisce, il parametro di lisciamento preso in considerazione conduce ad un eccessivo lisciamento nel caso di dati che non seguono una distribuzione normale.

Stimatori del nucleo adattivi

Come anticipato nel capitolo precedente, in questo elaborato si lavora su zone a differente concentrazione, poiché i fotoni emessi da sorgente sono più concentrati lungo la direzione di emissione. Pertanto, per ottenere maggiore precisione in fase di stima della densità, è opportuno ricorrere ad un parametro di lisciamento variabile.

Seguendo la regola di Abramson (1982) per ottenere un parametro di lisciamento variabile e dipendente dall'osservazione *i*-esima si ottiene

$$h^A_{i,NS} = h_{NS} \left[\tilde{f}_{h_{NS}}(\boldsymbol{x}_i)\right]^{-\frac{1}{2}}, \tag{2.9}$$

mentre secondo la regola di Silverman (1986) si ottiene:

$$h^S_{i,NS} = h_{NS} \left[\frac{1}{g}\tilde{f}_{h_{NS}}(\boldsymbol{x}_i)\right]^{-\beta}, \tag{2.10}$$

con $\tilde{f}_{h_{NS}}$ densità pilota ottenuta applicando il parametro di lisciamento in (2.6). Il valore di β comunemente utilizzato è 0.5, in quanto questa scelta comporta un miglior comportamento dello stimatore *kernel* sulle code della distribuzione.

2.3.3 Stima *kernel binned* per dati direzionali

Come anticipato nel Paragrafo 2.2, per far fronte alla complessità computazionale del problema in esame, viene utilizzata una stima *kernel binned*. Considerando l'elevata concentrazione dei fotoni e l'enorme mole di dati, lavorare con meno osservazioni renderà computazionalmente meno onerosa non solo la fase di stima della densità, ma anche la fase di individuazione delle componenti connesse.

L'estensione della stima *kernel binned* al caso sferico avviene seguendo l'espressione (1.14), utilizzando il nucleo von Mises-Fisher e la griglia triangolare descritta nel Paragrafo 2.2 (Figura 2.3). Il centroide di ogni triangolo costituirà il rappresentante della cella e

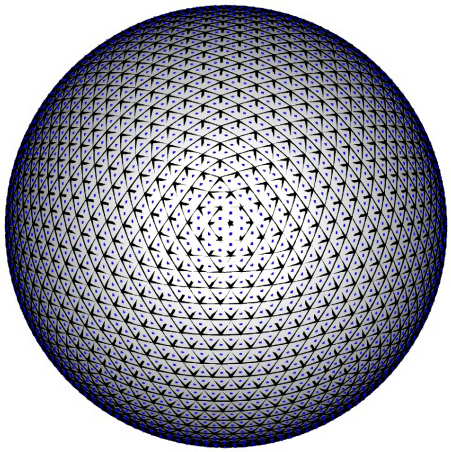

Figura 2.3: Tassellazione triangolare su una sfera di raggio unitario. I punti in blu rappresentano i centroidi di ogni cella.

il numero di osservazioni che cadono in quella determinata cella contribuiranno alla determinazione del peso relativo alla cella stessa.

Anche in questo contesto, la vicinanza dello stimatore *kernel* per dati direzionali allo stimatore *kernel* tradizionale, e quindi l'accuratezza della stima, dipende dalla dimensione della griglia. Si rimanda al Paragrafo 2.5 per approfondimenti.

2.4 Individuazione dei gruppi

Nel Paragrafo 1.3 è stato introdotto un metodo di *clustering* non parametrico basato sul concetto di curve di livello. La naturale estensione al contesto in esame comporta alcune variazioni rispetto agli esempi introdotti nel capitolo precedente.

Una volta determinata la stima *kernel* della densità mediante lo stimatore *binned* descritto nel paragrafo precedente, la procedura prevede l'identificazione delle regioni connesse associate alle curve di livello di tale stima. A questo scopo, entra nuovamente in gioco la griglia definita nel Paragrafo 2.2 e già utilizzata per la costruzione della stima *binned*. Tale griglia rappresenta lo strumento alla base per

la costruzione di un grafo $\mathscr{G}$ del quale si identificano le componenti connesse.

In particolare, a partire dal campione $(\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$ ripartito nelle nelle celle $\{g_j\}_{j\in\mathbb{Z}^d}$ della griglia, ogni cella rappresenta un nodo di $\mathscr{G}$ e due nodi risultano connessi attraverso un arco se condividono un lato della cella ad essi associata. L'algoritmo procede, come descritto nel Paragrafo 1.3, identificando, per una sequenza di valori di λ, le componenti connesse del sottografo $\mathscr{G}_\lambda$ indotto dall'insieme di livello

$$\mathscr{S}(\lambda) = \{g_j : j \in \mathbb{Z}^d, \tilde{f}_h(g_j) > \lambda\},$$

dove $\tilde{f}_h$ è la stima *binned* (1.14). In altre parole, si procede al raggruppamento delle celle della griglia in luogo delle osservazioni campionarie, che ereditano la classe di assegnazione della cella di cui fanno parte, producendo un notevole vantaggio computazionale.

Inoltre, dopo aver individuato i nuclei dei *cluster*, rimangono delle osservazioni non allocate, che in questo lavoro vengono assunte appartenenti al *background*, in quanto la concentrazione dei fotoni emessi da segnale è maggiore lungo la direzione di emissione. L'idea, dunque, si basa sul fatto che una moda (sorgente) è presente laddove è individuata una maggiore concentrazione di massa di probabilità rispetto al vicinato. Pertanto, viene identificata l'*excess mass*, ovvero la massa di probabilità sopra una certa soglia di densità (Muller e Sawitzki, 1991). Tale soglia viene fissata in corrispondenza del valore più alto che identifica una regione modale. La Figura 2.4 rappresenta un'illustrazione di questa idea sull'asse reale.

2.5 Aspetti computazionali

L'utilizzo della stima *kernel binned* e la successiva identificazione delle componenti connesse del grafo associato alla regola di *binning* devono essere supportati dalla ricerca di un giusto compromesso tra guadagno computazionale e distorsione della stima.

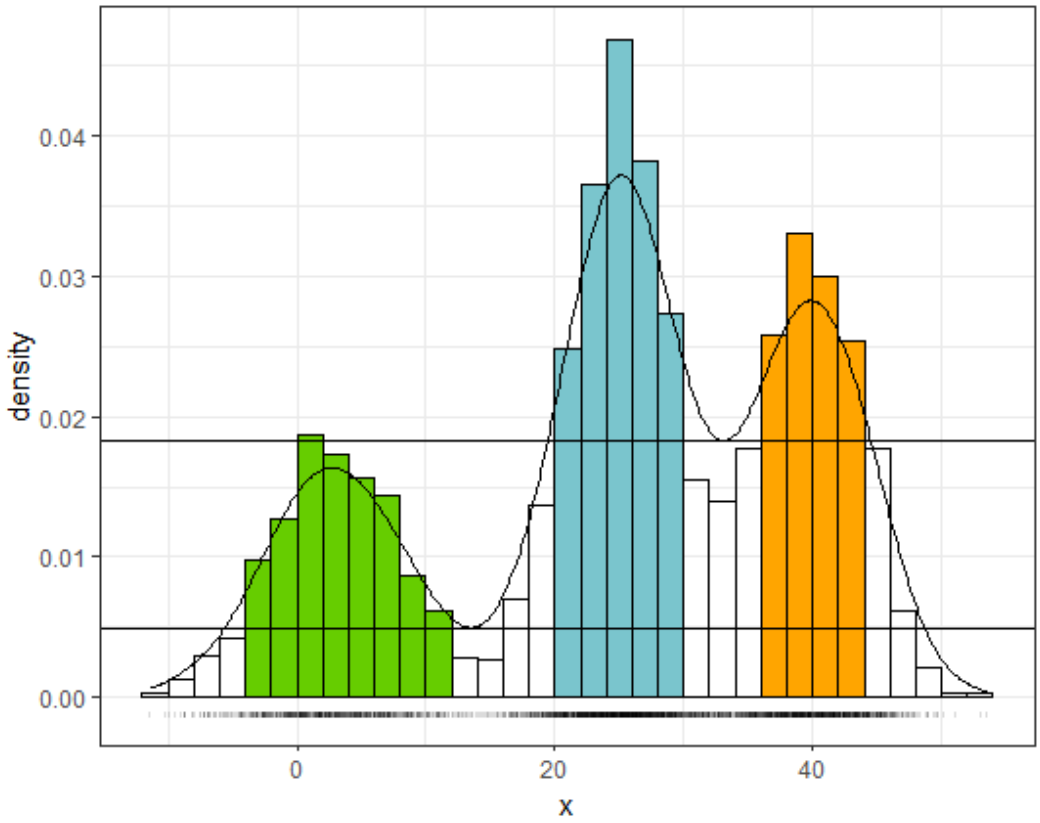

Figura 2.4: Funzione di densità e le relative zone a maggiore concentrazione di massa di probabilità.

Qualora venisse scelta una griglia poco fitta, si contrarrebbe il rischio che più sorgenti cadano in una stessa cella. Conseguentemente, laddove devessero essere presenti più mode, la stima *kernel binned* individuerebbe una singola moda, e, in ultimo, avremmo un unico *cluster*.

D'altro canto, se si scegliesse una griglia più fitta, il numero di celle rischierebbe di eccedere il numero di osservazioni, rendendo vano l'utilizzo della stima *kernel binned* per il superamento dei limiti computazionali. In questo secondo caso, però, il problema può essere arginato mediante un accorgimento che ha inoltre il vantaggio di ridurre ulteriormente l'onere computazionale della procedura.

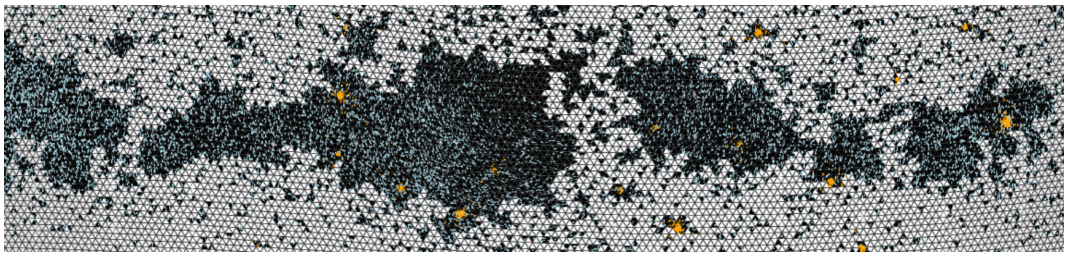

Figura 2.5: Porzione della sfera in cui vengono eliminate le celle vuote per poter lavorare su regioni indipendenti.

Le celle vuote potranno essere eliminate, in modo tale che sia possibile distinguere le regioni caratterizzate dalla presenza di osservazioni da regioni a bassa densità. In questo modo, sarà possibile lavorare su una griglia composta da regioni sconnesse ed applicare l'algoritmo di *clustering* alle varie componenti indipendenti l'una dall'altra, riducendo considerevolmente i tempi legati alla stima della densità e all'individuazione delle componenti connesse. Un esempio di suddivisione di una porzione di sfera in regioni indipendenti è presente in Figura 2.5.

Questo metodo presenta un ulteriore vantaggio rispetto alla suddivisione della sfera in sotto-regioni spaziali: sfruttando l'informazione data dalle celle vuote, le regioni saranno individuate riducendo il rischio di separare fotoni relativi a una stessa sorgente, in quanto circondate da porzioni di sfera in cui non cadono osservazioni.

Capitolo 3

Analisi empirica

3.1 Descrizione dei dati

L'analisi che segue si prefigge l'obiettivo di applicare la metodologia sviluppata nei capitoli precedenti ad un insieme di dati simulati, in un contesto in cui le emissioni provengono sia da processi di segnale sia dal rumore di fondo, al fine di separare le sorgenti dal *background* diffuso.

Le unità statistiche sono i fotoni e le informazioni a disposizione[1], raccolte dal telescopio Fermi con il rilevatore LAT, vengono registrate a seguito di una ricostruzione delle variabili utili all'analisi tramite un algoritmo (Ackermann *et al.*, 2012). Ogni fotone emesso è identificato dal rilevatore LAT e l'evento viene ricostruito grazie al tracciatore e al calorimetro. Il primo ricostruisce la direzione di ogni evento, mentre il secondo ricostruisce l'energia del fotone. Dunque, oltre alla posizione, che rappresenta la direzione dell'emissione, vengono registrati l'energia dell'evento, l'angolo di incidenza, il tempo della rilevazione e il tipo di evento che caratterizza la qualità della rilevazione. Quest'ultima è affetta da errore di misura associato alle componenti del LAT ed è espressa tramite la *point spread function* (Ackermann *et al.*, 2013).

[1] `https://fermi.gsfc.nasa.gov/ssc/data/analysis/documentation/Cicerone/Cicerone_Data/LAT_Data_Columns.html`

In questo lavoro, in un primo momento si utilizza esclusivamente l'informazione legata alla direzione dell'emissione, che viene espressa in coordinate galattiche. In questo modo il centro del sistema di coordinate cartesiane è il punto (0°, 0°), in cui si colloca il buco nero supermassiccio *Sagittarius A** (Cafardo e Nemmen, 2021).

La procedura descritta nei capitoli precedenti è stata valutata su un insieme di dati, costruito simulando tutte le sorgenti con più di tre fotoni registrate dal 3FHL, il terzo catalogo di sorgenti rilevate dal Fermi LAT con energia superiore ai 10 GeV durante i primi sette anni di rilevazione (Ajello e Atwood, 2017). L'applicazione delle metodologie presentate è stata estesa a tutta la sfera celeste, comprendendo anche il *background* diffuso. Ai livelli di energia considerati, infatti, la componente isotropica è da considerarsi trascurabile. Si considerano le emissioni di fotoni rilevati a partire dal lancio del telescopio Fermi a giugno del 2008 e per un periodo di tempo lungo circa 7.2 anni. La simulazione è stata effettuata considerando un modello di Poisson, con specificazione diversa per fotoni provenienti da sorgenti[2] e fotoni provenienti dal *background*[3]. Si veda, per maggiori dettagli, Costantin *et al.* (2020).

La distribuzione dei dati è alquanto eterogenea, con circa l'84% proveniente dal rumore di fondo. In particolare, si dispone di 396466 eventi simulati dal *background* diffuso, con una maggiore concentrazione a latitudine 0°, che decresce allontanandosi dal piano galattico. Inoltre, 73318 fotoni sono emessi da 1529 sorgenti, la cui grandezza varia da 4 a 3572 fotoni (Figura 3.1). La sorgente che emette il maggior numero di fotoni, 3572, è la *Pulsar delle Vele* (PSR B0833-45 o PSR J0835-4510) e giace sul piano galattico. L'energia dei fotoni rilevata dal LAT è misurata in mega elettronvolt (MeV). I dati simulati dal *background* diffuso hanno energia compresa tra 10000.01 MeV e 762398.19 MeV, mentre quelli simulati da sorgenti tra 10000.18 MeV

[2] `https://fermi.gsfc.nasa.gov/ssc/data/analysis/scitools/source_models.html`

[3] `https://fermi.gsfc.nasa.gov/ssc/data/access/lat/BackgroundModels.html`

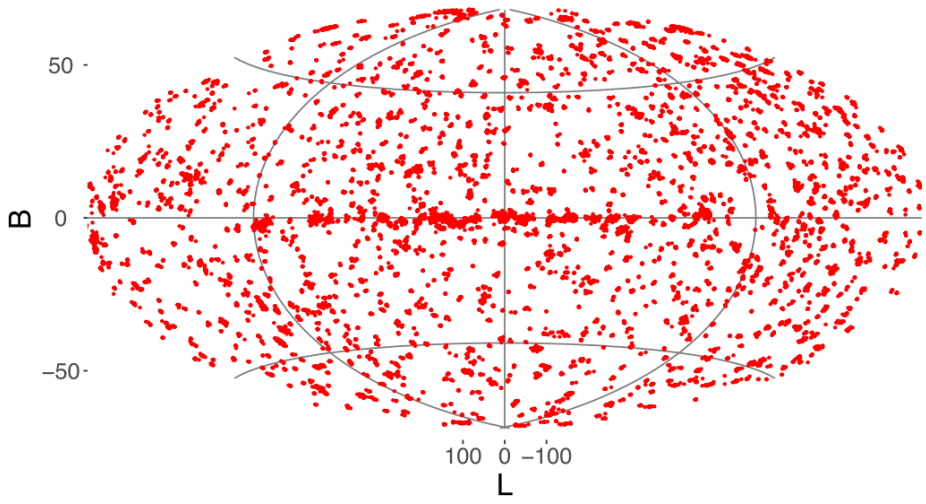

Figura 3.1: Proiezione di Aitoff delle direzioni dei fotoni simulati da sorgenti sull'intera sfera celeste in circa 7.2 anni.

e 999012.38 MeV. Come si vede in Figura 3.2 la maggior parte degli eventi presenta energia che si colloca alla base della scala di valori considerata.

3.2 Implementazione della procedura

Il primo passo dell'analisi consiste nella localizzazione dei punti e nella costruzione della griglia secondo le modalità descritte nel Paragrafo 2.2.

Data la natura dei dati, è stato utile inoltre considerare alcuni ulteriori accorgimenti.

Idealmente, sarebbe opportuno utilizzare una griglia sferica più fitta lungo il piano galattico e le cui celle sono più grandi altrove, così da rendere possibile l'identificazione dei *cluster* laddove è presente una maggiore concentrazione di fotoni ed il rischio che sorgenti vicine non vengano distinte è maggiore. Ciò non è realizzabile in base a come viene costruita la griglia, ma è possibile unire più griglie per realizza-

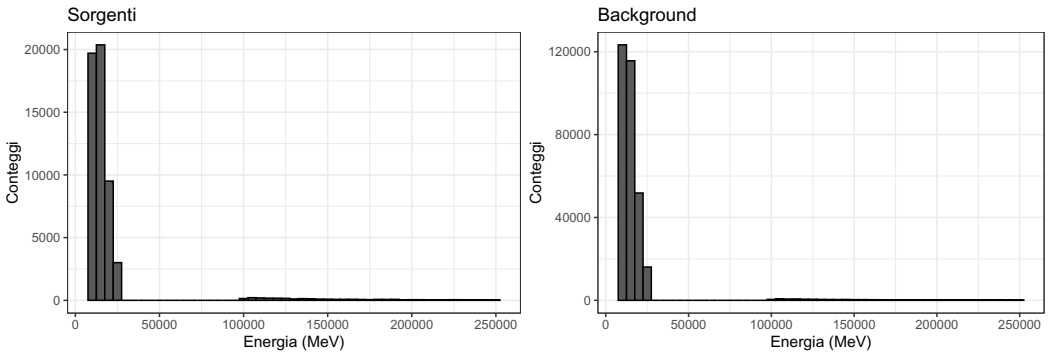

Figura 3.2: Istogramma dell'energia degli eventi simulati dal sorgenti (a sinistra), e degli eventi simulati da *background* (a destra).

re quanto spiegato sopra. Se si costruisse una griglia sufficientemente fitta per i poli, ma non abbastanza per il piano galattico, eliminando le celle vuote che separano regioni occupate da fotoni, si individuerebbe un'unica componente non scomponibile presso la zona dell'equatore, perché non ci sarebbero celle vuote tra regioni occupate. Per questo, una volta individuata questa componente più grande, si posizionano le relative osservazioni su una griglia più fitta: in questo modo diventa più semplice individuare celle vuote, che, una volta eliminate, permetteranno l'individuazione di regioni sconnesse. Nonostante si utilizzi una griglia più fitta per la regione centrale, con un conseguente aumento del numero di celle, che apparentemente renderebbe la stima della densità e l'individuazione dei *cluster* più onerose, il vantaggio computazionale non viene compromesso, perché la determinazione di più porzioni sconnesse nella zona centrale è permessa proprio da questa modifica.

Per poter suddividere ulteriormente la regione centrale, si eliminano, oltre alle celle vuote, quelle contenenti una o due osservazioni, che vengono assegnate al *background*.

Come mostrato dalla Figura 3.3, si individuano in questo modo tre regioni principali: una ai poli, una nella zona presso l'Equatore e una componente connessa più grande. Le prime due, al contrario della

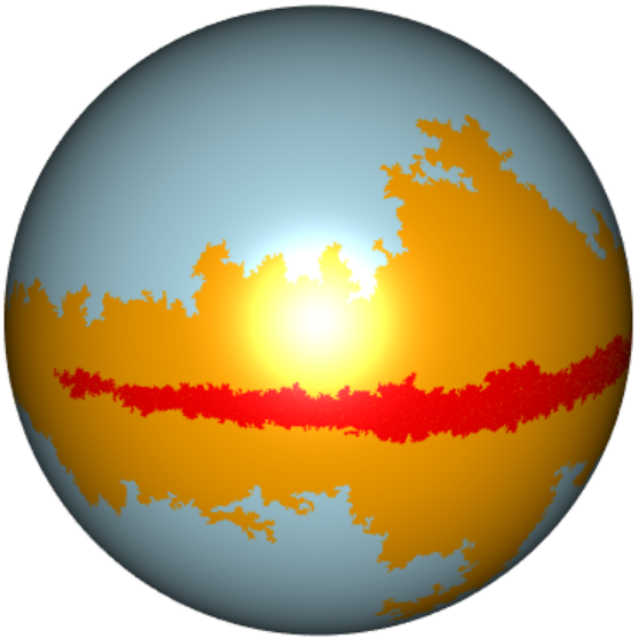

Figura 3.3: Partizione della griglia in tre zone: poli (azzurro), zona centrale (arancione) e componente connessa più grande (rosso).

terza, sono scomponibili ulteriormente in porzioni indipendenti.

Una volta individuata la partizione della griglia, si stima la densità e si applica la procedura di *clustering* descritta sulle singole porzioni di griglia sconnesse.

Per quanto concerne la scelta del parametro di lisciamento, dopo aver effettuato delle prove in varie porzioni della sfera appartenenti ai poli, alla regione centrale e alla componente connessa di dimensione maggiore, è stata individuata una regola del pollice empirica. Seguendo la (2.10), si utilizza un parametro di lisciamento variabile, ma, rispetto a quanto esposto nel capitolo precedente, la densità pilota è ottenuta applicando il parametro di lisciamento $h_{NS}/10$. Questa regola è stata utilizzata in tutte le porzioni della sfera.

In realtà, a seguito della eliminazione di celle vuote, molte regioni sono composte solamente da una o due celle. Quando il numero

di osservazioni appartenenti a queste regioni è inferiore ad una soglia, in questo lavoro posta pari a 5, si sceglie di considerarle come appartenenti al *background*. In caso contrario, vengono considerate come fotoni emessi da sorgenti, in quanto rappresentano picchi della densità sottostante ai dati, per cui non è necessario effettuare il raggruppamento.

3.3 Valutazione della procedura

Dal momento che i dati sono tratti da un catalogo di sorgenti già rilevate, possiamo valutare la bontà della procedura grazie al fatto che la sorgente emittente di ogni fotone è nota.

Per valutare la capacità della procedura di distinguere tra segnale e *background*, vengono utilizzati il *True Positive Rate* (TPR) e il *False Positive Rate* (FPR). Il primo indice è definito come la proporzione di vere sorgenti correttamente rilevate, mentre il secondo corrisponde alla proporzione di componenti stimate che non sono in realtà associate ad alcuna sorgente.

Per quanto concerne la qualità dell'assegnazione dei singoli fotoni, verrà utilizzato la *Misclassification Error Distance* (MED), inizialmente introdotta da Régnier (1983) come distanza tra partizioni di un insieme finito. Un modo comune per esprimere la discrepanza tra due insiemi C e D consiste nel quantificare il contenuto della loro differenza simmetrica, $C \triangle D = (C \cup D) \setminus (C \cap D)$. Questa definizione naturale di distanza tra insiemi può essere estesa per definire la distanza tra due *clustering* $\mathscr{C} = \{C_1, ..., C_r\}$ e $\mathscr{D} = \{D_1, ..., D_s\}$, aggiungendo i contributi delle regioni che i loro *cluster* più simili non hanno in comune:

$$d_M(\mathscr{C}, \mathscr{D}) = \frac{1}{2} \min_{\sigma \in \mathscr{P}_s} \sum_{i}^{s} P(C_i \triangle D_{\sigma(i)}), \tag{3.1}$$

con $\mathscr{P}_s$ insieme di tutte le permutazioni di s elementi e $r \leq s$. Intuitivamente, $d_M(\mathscr{C}, \mathscr{D})$ rappresenta la minima massa di probabilità che

deve essere spostata per trasformare $\mathscr{C}$ in $\mathscr{D}$ o viceversa. La versione empirica della (3.1) coincide con il MED, ovvero $MED(\mathscr{C}, \mathscr{D}) = \hat{d}(\mathscr{C}, \mathscr{D})$, che, pertanto, rappresenta la proporzione minima di punti che dovrebbero essere riassegnati in modo tale che $\mathscr{C}$ e $\mathscr{D}$ coincidano. Nonostante l'apparente difficoltà della ricerca di tutte le possibili permutazioni di s elementi, questa può essere superata grazie agli algoritmi che risolvono problemi di assegnazione.

L'analisi è stata svolta tramite il linguaggio di programmazione R (R Core Team, 2022). In particolare, per la costruzione della griglia è stato utilizzato il pacchetto `icosa` (Kocsis, 2021), per la sua suddivisione in regioni indipendenti il pacchetto `igraph` (Csardi e Nepusz, 2006), per la selezione del parametro di lisciamento si è fatto uso di `ks` (Duong, 2022), per la ricerca delle componenti connesse è stato utilizzato il pacchetto `pdfCluster` (Azzalini e Menardi, 2014), mentre il calcolo del MED è stato effettuato mediante l'ausilio della libreria `clue` (Hornik, 2005).

3.4 Discussione dei risultati

I risultati della procedura descritta sono riportati nella Tabella 3.1 e mostrano una buona *perfomance* generale riguardo alla rilevazione delle sorgenti. L'algoritmo classifica correttamente oltre l'80% dei fotoni e identifica quasi il 90% delle sorgenti. Tuttavia, una parte non indifferente di osservazioni provenienti da *background* viene erroneamente classificata come segnale.

Questo avviene principalmente a causa del metodo utilizzato per la partizione della griglia: nonostante le osservazioni appartenenti al *background* classificate come sorgenti costituiscano il 21.7% delle osservazioni appartenenti al rumore di fondo, vengono individuati molti *cluster* spuri perché sono presenti molte regioni contenenti esclusivamente fotoni emessi dal *background*. Questi vengono erroneamente classificati come sorgenti, con un conseguente aumento del FPR.

	rumore	sorgente
$\widehat{rumore}$	310430	7239
$\widehat{sorgente}$	86036	66079
MED	0.1985	
TPR	0.8947	
FPR	0.8515	

Tabella 3.1: Matrice di confusione relativa a tutte le osservazioni sulla superficie sferica e risultati del raggruppamento.

Si individuano due ulteriori motivi per cui il numero di errate classificazioni aumenta.

- Quando la stima della densità in una regione è unimodale, tutte le osservazioni appartenenti alla data porzione di sfera vengono classificate come segnale, perché l'albero dei *cluster* è composto da una sola foglia. Questo vale non solo nel caso in cui nella regione è presente una singola sorgente, ma anche quando la porzione contiene solo fotoni emessi dal *background*, in quanto l'assenza di gruppi, nel *clustering* modale, corrisponde concettualmente all'identificazione di una densità unimodale.

- Quando nella regione sono presenti più sorgenti distanti l'una dall'altra, ovvero nel caso in cui le mode della stima della densità sono separate da una regione a densità molto bassa, i relativi *cluster cores* individuati risultano eccessivamente estesi e al gruppo vengono assegnate molte osservazioni relative al *background*.

Esaminando in dettaglio i risultati ottenuti dal raggruppamento nelle diverse sezioni rappresentate in Figura 3.3, le conclusioni non cambiano. Tuttavia, dalla Tabella 3.2b, si nota un aumento del FPR nella regione centrale rispetto alle altre sezioni. In questa porzione, infatti, data l'elevata mole di dati, si individuano molte regioni composte esclusivamente da fotoni emessi da rumore di fondo, con un conseguente aumento dei *cluster* spuri e delle errate classificazioni.

	rumore	sorgente
$\widehat{rumore}$	48771	1444
$\widehat{sorgente}$	24819	28722
MED	0.2531	
TPR	0.9260	
FPR	0.7930	

(a) Poli.

	rumore	sorgente
$\widehat{rumore}$	146399	1962
$\widehat{sorgente}$	40294	21718
MED	0.2009	
TPR	0.9234	
FPR	0.9122	

(b) Regione centrale.

	rumore	sorgente
$\widehat{rumore}$	115260	3833
$\widehat{sorgente}$	20923	15639
MED	0.1590	
TPR	0.3446	
FPR	0.6015	

(c) Componente connessa più grande.

Tabella 3.2: Matrici di confusione e risultati del raggruppamento in base alla zona considerata.

Dalla Tabella 3.2c, invece, si nota una diminuzione del TPR a causa della complessità del raggruppamento nella regione considerata e dell'inadeguatezza del parametro di lisciamento, ma, d'altra parte, si avverte una diminuzione sia del MED che del FPR. In questo caso, infatti, l'algoritmo di *clustering* viene applicato all'intera regione e, di conseguenza, non si presentano i problemi esposti sopra.

É noto che il numero minimo di fotoni emessi da ciascuna sorgente è pari a quattro. Se si eliminassero i gruppi composti da meno di quattro fotoni, che costituiscono il 16.2% delle sorgenti stimate, si avrebbe una diminuzione sia del FPR sia del MED, in particolare nelle regioni ai poli, ma rimarrebbero comunque i problemi precedentemente elencati.

3.5 Possibili avanzamenti

Per ridurre il numero di osservazioni misclassificate appartenenti al rumore di fondo, un possibile modo di procedere consiste nel valutare separatamente la stima della densità delle osservazioni su cui è stato effettuato il raggruppamento e una stima della densità relativa ai soli dati di *background* su tutta la sfera, ottenuta nelle stesse condizioni, quindi a parità di parametro di lisciamento e con la stessa partizione della griglia.

	rumore	sorgente
$\widehat{rumore}$	219631	6932
$\widehat{sorgente}$	57895	66386
MED	0.1847	
TPR	0.8940	
FPR	0.7777	

Tabella 3.3: Matrice di confusione relativa a tutte le osservazioni sulla superficie sferica e risultati del raggruppamento.

	rumore	sorgente
$\widehat{rumore}$	49249	1809
$\widehat{sorgente}$	19578	32780
MED	0.2068	
TPR	0.9290	
FPR	0.6682	

(a) Poli.

	rumore	sorgente
$\widehat{rumore}$	104371	1937
$\widehat{sorgente}$	28835	22653
MED	0.1950	
TPR	0.8711	
FPR	0.9037	

(b) Regione centrale.

	rumore	sorgente
$\widehat{rumore}$	66011	3186
$\widehat{sorgente}$	9482	10953
MED	0.1413	
TPR	0.42	
FPR	0.475	

(c) Componente connessa più grande.

Tabella 3.4: Matrici di confusione e risultati del raggruppamento in base alla zona considerata.

A questo scopo, è possibile sfruttare l'informazione relativa al tempo di emissione dei fotoni. In particolare, è possibile suddividere il dataset relativo ai fotoni emessi dal *background* in due porzioni: il primo 30% è costituito dalle rilevazioni relative al periodo che intercorre tra 247001879 MET e 318075354 MET, corrispondente circa ai primi due anni e mezzo di rilevazione, mentre il rimanente 70% è relativo ai fotoni osservati da 318079403 MET a 474173562 MET. Considerando questa seconda porzione e i fotoni emessi da sorgente, è necessario ricalcolare la partizione della griglia relativa a queste osservazioni, secondo le modalità descritte nei capitoli precedenti, in modo tale che l'algoritmo di *clustering* possa essere applicato a queste rilevazioni. Le Tabelle 3.3, 3.4c, 3.4a, 3.4b riportano i risultati del raggruppamento su questo insieme di stima ridotto e sono confrontabili con i precedenti.

	rumore	sorgente
$\widehat{rumore}$	236104	14170
$\widehat{sorgente}$	41422	59148
MED	0.1584	
TPR	0.8777	
FPR	0.7808	

Tabella 3.5: Matrice di confusione relativa a tutte le osservazioni sulla superficie sferica e risultati del raggruppamento dopo la correzione.

	rumore	sorgente
$\widehat{rumore}$	56048	7191
$\widehat{sorgente}$	12779	27398
MED	0.1931	
TPR	0.9097	
FPR	0.6739	

(a) Poli.

	rumore	sorgente
$\widehat{rumore}$	114045	3793
$\widehat{sorgente}$	19161	20797
MED	0.1454	
TPR	0.8396	
FPR	0.9067	

(b) Regione centrale.

	rumore	sorgente
$\widehat{rumore}$	68418	3331
$\widehat{sorgente}$	7075	10808
MED	0.1160	
TPR	0.4	
FPR	0.2	

(c) Componente connessa più grande.

Tabella 3.6: Matrici di confusione e risultati del raggruppamento dopo la correzione in base alla zona considerata.

Una volta stimate le densità, si considera la loro differenza e si assegnano al *background* i fotoni classificati come sorgente e in corrispondenza dei quali la differenza fra le densità stimate non risulta maggiore di zero.

I risultati riportati dalle Tabelle 3.5, 3.6c, 3.6a, 3.6b, se confrontati con i risulati ottenuti prima della correzione, mostrano un generale miglioramento in termini di MED. Questo metodo, infatti, permette di evitare l'estensione dei nuclei dei *cluster*, e, di conseguenza, la misclassificazione di molte osservazioni appartenenti al *background*. Tuttavia, non si avverte un miglioramento in termini di FPR, ma ciò è dovuto al modo in cui è stata calcolata la partizione della griglia. Le osservazioni relative al primo 30% di *background*, infatti, nella maggior parte dei casi non appartengono alle regioni indipendenti individuate e su cui viene effettuato il raggruppamento. Pertanto, la

differenza fra le densità stimate nelle regioni indipendenti risulta positiva e non vengono eliminati *cluster* spuri.

Considerando la Tabella 3.6c, invece, relativa alla regione più grande, l'unica zona non ulteriormente suddivisa, si ottiene un sensibile miglioramento sia in termini di MED, sia in termini di FPR, che decresce del 27.5%, a fronte di una diminuzione del TPR del 2%. Dunque, questi risultati indicano la correttezza del metodo di correzione proposto, ma evidenziano l'inadeguatezza del modo in cui è stata suddivisa la griglia.

Conclusioni

In questo elaborato ho approfondito il tema del *clustering* non parametrico con l'obiettivo di analizzare i dati sui fotoni emessi dalle sorgenti celesti raccolti dal telescopio spaziale Fermi LAT, proponendo una soluzione che cercasse di separare i fotoni emessi da sorgenti da quelli emessi da *background* e che, al contempo, superasse le difficoltà che emergono dal punto di vista computazionale.

Nel Capitolo 1, ho approfondito, dal punto di vista teorico, le tecniche utili per il conseguimento dell'obiettivo, riservando particolare attenzione ai metodi comunemente adottati per la selezione del parametro di lisciamento e alla variante *binned* della stima *kernel*.

Le tecniche introdotte sono state successivamente riprese nel Capitolo 2 e riadattate al contesto in esame. Si è posto particolare riguardo agli aspetti computazionali del problema, per cui è stata proposta una soluzione che permettesse di lavorare sull'intera sfera e sfruttasse la stima *kernel binned*, la costruzione e la partizione della griglia per ridurre drasticamente i tempi impiegati dall'algoritmo.

Nel Capitolo 3 sono stati presentati i risultati dell'applicazione della metodologia proposta a dati simulati da sorgenti e dal *background* diffuso. Ciò ha portato a buoni risultati per quanto riguarda la rilevazione delle reali sorgenti. Tuttavia, è necessario compiere miglioramenti dal punto di vista della discriminazione del segnale dal rumore di fondo, cercando di ridurre il tasso di falsi positivi rilevati.

L'argomento fornisce spunti di approfondimento: in particolare, sviluppi futuri si potranno incentrare sulla ricerca di metodi alternativi per la partizione della griglia, in modo tale da evitare la creazione

di un numero eccessivo di regioni indipendenti, con una conseguente diminuzione del numero di *cluster* spuri individuati. Ricerche future, inoltre, possono incentrarsi sulla scelta del parametro di lisciamento iniziale e sulla ricerca di un modo per attribuire una misura di significatività alle sorgenti rilevate. A seguito dell'applicazione della procedura di *clustering*, è possibile escludere eventuali gruppi qualora non risultino significativi. La letteratura relativa a questo ambito è piuttosto scarsa e si occupa in particolar modo di valutare la multimodalità della funzione di densità. Questo obiettivo può essere perseguito riferendosi al concetto di *excess mass*, ovvero valutando la massa di probabilità al di sopra di una certa soglia di densità. Qualora questa massa risultasse non significativa, le osservazioni appartenenti al *cluster* verrebbero dunque classificate come *background*.

Bibliografia

Abramson I. S. (1982). On bandwidth variation in kernel estimates: a square root law. *The annals of Statistics*, pp. 1217–1223.

Ackermann M.; Ajello M.; Albert A.; Allafort A. (2012). The fermi large area telescope on orbit: Event classification, instrument response functions, and calibration. *The Astrophysical Journal Supplement Series*, **203**(1), 4.

Ackermann M.; Ajello M.; Allafort A. (2013). Determination of the point-spread function for the fermi large area telescope from on-orbit data and limits on pair halos of active galactic nuclei. *The Astrophysical Journal*, **765**(1), 54.

Ajello M.; Atwood W. B. (2017). 3fhl: The third catalog of hard fermi-lat sources. *The Astrophysical Journal Supplement Series*, **232**(2), 18.

Azzalini A.; Menardi G. (2014). Clustering via nonparametric density estimation: The R package pdfCluster. *Journal of Statistical Software*, **57**(11), 1–26.

Azzalini A.; Torelli N. (2007). Clustering via nonparametric density estimation. *Statistics and Computing*, **17**, 71–80.

Bai Z.; Rao C.; Zhao L. (1988). Kernel estimators of density function of directional data. *Journal of Multivariate Analysis*, **27**(1), 24–39.

Bouveyron C.; Celeux G.; Murphy T. B.; Raftery A. E. (2019). *Bibliography*, p. 386–414. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Bowman A. W.; Azzalini A. (1999). Applied smoothing techniques for data analysis : the kernel approach with s-plus illustrations. *Journal of the American Statistical Association*, **94**, 982.

Breiman L.; Meisel W.; Purcell E. (1977). Variable kernel estimates of multivariate densities. *Technometrics*, **19**(2), 135–144.

Cafardo F.; Nemmen R. (2021). Fermi-LAT Observations of Sagittarius A^* Imaging Analysis. *The Astrophysical Journal*, **918**(1), 30.

Carmichael J. W.; Julius R. S. (1968). Finding natural clusters. *Systematic Biology*, **17**, 144–150.

Chacón J. E. (2015). A population background for nonparametric density-based clustering. *Statistical Science*, **30**(4).

Chacón J.; Duong T. (2018). *Multivariate Kernel Smoothing and its Applications*.

Costantin D.; Menardi G.; Brazzale A.; Bastieri D.; Fan J. (2020). A novel approach for pre-filtering event sources using the von mises–fisher distribution. *Astrophysics and Space Science*, **365**.

Csardi G.; Nepusz T. (2006). The igraph software package for complex network research. *InterJournal*, **Complex Systems**, 1695.

Deheuvels P. (1977). Estimation non paramétrique de la densité par histogrammes généralisés. *Revue de Statistique Appliquée*, **25**(3), 5–42.

Duong T. (2022). *ks: Kernel Smoothing*. R package version 1.14.0.

Fukunaga K.; Hostetler L. D. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. Inf. Theory*, **21**, 32–40.

García-Portugués E. (2013). Exact risk improvement of bandwidth selectors for kernel density estimation with directional data. *Electronic Journal of Statistics*, **7**, 1655–1685.

Hall P.; Wand M. P. (1994). On the accuracy of binned kernel density estimators. *Econometrics eJournal*.

Hall P.; Watson G. S.; Cabrera J. (1987). Kernel density estimation with spherical data. *Biometrika*, **74**(4), 751–762.

Hartigan J. A. (1975). *Clustering Algorithms*. John Wiley & Sons, Inc., USA, 99th edizione.

Hobson M. (2009). *Bayesian Methods in Cosmology*. Cambridge University Press.

Hornik K. (2005). A clue for cluster ensembles. *Journal of Statistical Software*, **14**(12), 1–25.

Kaufman L.; Rousseeuw P. J. (2005). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, New York.

Kocsis A. T. (2021). *icosa: Global Triangular and Penta-Hexagonal Grids Based on Tessellated Icosahedra*. R package version 0.10.1.

Mardia K.; Jupp P. (2000). *Directional Statistics*, volume 494. John Wiley & Sons.

Marron J. S.; Wand M. P. (1992). Exact mean integrated squared error. *The Annals of Statistics*, **20**(2), 712–736.

Mattox J.; Bertsch D.; Chiang J.; Dingus B.; Digel S.; Esposito J.; Fierro J.; Hartman R.; Hunter S.; Kanbach G.; Kniffen D.; Lin C.;

Macomb D.; Mayer-Hasselwander H.; Michelson P.; Montigny C.; Mukherjee R.; Nolan P.; Ramanamurthy P.; Willis T. (1996). The likelihood analysis of egret data. *The Astrophysical Journal*, **461**, 396.

Menardi G. (2016). A Review on Modal Clustering. *International Statistical Review*, **84**(3), 413–433.

Menardi G.; Azzalini A. (2014). An advancement in clustering via nonparametric density estimation. *Statistics and Computing*, **24**, 753–767.

Montin A. (2021). Individuazione di sorgenti di raggi gamma mediante clustering non parametrico: un'analisi dei dati fermi LAT. `https://hdl.handle.net/20.500.12608/2085`. Università degli Studi di Padova.

Montin A.; Brazzale A. R.; Menardi G. (2023). Locating γ-ray sources on the celestial sphere via modal clustering.

Muller D. W.; Sawitzki G. (1991). Excess mass estimates and tests for multimodality. *Journal of the American Statistical Association*, **86**(415), 738–746.

Parzen E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, **33**(3), 1065–1076.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Régnier S. (1983). Sur quelques aspects mathématiques des problèmes de classification automatique. *Mathématiques et Sciences Humaines*, **82**, 13–29.

Scott D.; Sheather S. (1985). Kernel density estimation with binned data. *Communications in Statistics-theory and Methods*, **14**, 1353–1359.

Silverman B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London.

Stuetzle W. (2003). Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample. *J. Classification*, **20**, 025–047.

Taylor C. (2008). Automatic bandwidth selection for circular density estimation. *Computational Statistics & Data Analysis*, **52**, 3493–3500.

Terrell G. R.; Scott D. W. (1992). Variable kernel density estimation. *The Annals of Statistics*, **20**(3), 1236–1265.

Wand M.; Jones M. (1994). *Kernel smoothing*. Chapman and Hall/CRC.

Wenninger M. (1979). *The Geometrical Foundation of Natural Structure: A source book of Design.* Cambridge University Press.

Williams R. (1979). *Spherical Models*. Dover Pubns.